The Regulatory Potential of Marine Cyanobacteria: Transcriptional Factors and Small RNAs

studied in a comparative genomics approach

DISSERTATION

zur Erlangung des akademischen Grades
doctor rerum naturalium
(Dr. rer. nat.)
im Fach Biologie

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät I
Humboldt-Universität zu Berlin

von
Frau Diplom-Ingenieurin Ilka Maria Axmann
geboren am 21.12.1976 in Räckelwitz

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Christoph Markschies

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I:
Prof. Thomas Buckhout, PhD

Gutachter:

1. Prof. Dr. Thomas Börner
2. Prof. Dr. Karl Forchhammer
3. Dr. David J. Scanlan

Tag der mündlichen Prüfung: 17. Juli 2006

Abstract

Life on Earth is driven by the power of oxygenic photosynthesis transforming solar into chemical energy. Cyanobacteria such as *Prochlorococcus* and *Synechococcus* belong to the most important primary producers within the oceans and increasingly serve as models for photosynthetic organisms. To better understand the regulatory mechanisms in these picocyanobacteria, here the information from four genomes of closely related and even so ecologically divergent marine strains was used in a combined computational and experimental approach. Sequence signals and RNA-coding genes as novel elements in the regulation of gene expression were identified and their distribution along the phylogenetic gradient compared. Phylogenetic footprinting revealed a minimal conserved set of putative transcription factors, their binding sites and regulons. Sites for NtcA, LexA and ArsR-like regulators were found as well as new *cis* elements. RACE experiments verified several of these predicted sites belonging to the promoter region. The cyanobacterial core promoter shows similar features as known from *E. coli*. In addition, sequence elements at position -60 were frequently found associated with putative bidirectional promoters. For cyanophage P-SSP7, promoter studies revealed a conserved site, which might be recognised by the phage RNA polymerase or/and the host's RNA processing machinery. A search, focussing on conserved secondary structures, detected several non-coding RNAs named Yfr for cYanobacterial Functional RNA. One of these ncRNAs, Yfr1, appears dispensable for growth at greater depths. A highly conserved secondary structure and an unpaired CA dinucleotide repeat might represent essential functional elements of Yfr1. A comparative analysis of Yfr7 structures, transcript types and accumulation throughout the cyanobacterial radiation indicated this RNA as the likely homologue of the *E. coli* 6S RNA. Two distinct Yfr7 transcripts with a circadian but time-shifted expression pattern suggested a coupling of their expression to the circadian rhythm or light intensity. Experiments in *Synechocystis* discovered a novel antisense RNA-mediated regulatory mechanism that controls *isiA* mRNA abundance and assembly of IsiA-photosystem I supercomplexes. Functional assignments of these new elements in the future will contribute to a deeper understanding of the regulatory network of marine cyanobacteria and promote new studies on bacterial ncRNAs.

Keywords:
cyanobacteria, comparative genomics, non-coding RNAs, promoter

Zusammenfassung

Das Leben auf der Erde wird maßgeblich durch die Kraft der oxygenen Photosythese bestimmt, die Sonnen- in chemische Energie umwandelt. Cyanobakterien wie *Prochloro-* und *Synechococcus* zählen zu den wichtigsten primären Produzenten der Ozeane und werden zunehmend als Modelle für photosynthetische Organismen genutzt. Um die Regulationsmechanismen dieser Picocyanobakterien besser zu verstehen, wurde hier die Information von vier Genomen hochgradig verwandter aber dennoch ökologisch unterschiedlich angepasster mariner Stämme genutzt in einer Kombination aus computer-gestützten und experimentellen Untersuchungen. Sequenzsignale und RNA-kodierende Gene wurden als neuartige Regulationselemente identifiziert und entlang des phylogenetischen Gradienten verglichen. Mittels 'phylogenetic footprinting' konnte ein minimales, konserviertes Set möglicher Transkriptionsfaktoren, deren Bindestellen und Regulons aufgedeckt werden. NtcA-, LexA- und ArsR-ähnliche Motive wurden ebenso gefunden wie neue regulatorische Elemente. Mit Hilfe von RACE Experimenten wurden einige der vorhergesagten Bindestellen Promotorregionen zugeordnet. Der cyanobakterielle Promotor besitzt *E. coli* ähnliche Merkmale. Konservierte Sequenzelemente an Position -60 stellen vermutlich ein Merkmal bestimmter bidirektionaler Promotoren dar. Für den Cyanophagen P-SSP7 ergaben Promotorstudien eine konservierte Sequenz, die entweder durch die Phagen-RNA-Polymerase oder/und die Wirt's RNA Prozessierungsmaschinerie erkannt wird. Eine Suche nach konservierten Sekundärstrukturen detektierte mehrere nicht-kodierende RNAs, benannt Yfr für cYanobacterial Functional RNA. Eine dieser ncRNAs, Yfr1, scheint für das Wachstum in großen Tiefen verzichtbar. Eine konservierte Sekundärstruktur und eine ungepaarte CA-Dinukleotid-Wiederholung könnten funktionale Elemente von Yfr1 darstellen. Eine vergleichende Analyse der Strukturen, Transkriptvarianten und Akkumulation von Yfr7 innerhalb der cyanobakteriellen Linie ergab, dass diese RNA wahrscheinlich ein Homolog der *E. coli* 6S RNA ist. Zwei verschiedene Yfr7 Transkripte mit einem zirkadianen aber zeitversetzten Akkumulationsmuster lassen eine Verknüpfung ihrer Expression mit dem zirkadianen Rhythmus oder der Lichtintensität vermuten. Experimente in *Synechocystis* deckten einen neuartigen Regulationsmechanismus durch eine antisense RNA auf, welche die Menge der *isiA* mRNA kontrolliert und die Assemblierung von IsiA-Superkomplexen beeinflusst. Die funktionelle Zuordnung dieser neuen Elemente wird zu einem besseren Verständnis regulatorischer Netzwerke in marinen Cyanobakterien und darüber hinaus führen.

Schlagwörter:
Cyanobakterien, Genomvergleich, nicht-kodierende RNAs, Promotor

Phantasie ist wichtiger als Wissen, denn Wissen ist begrenzt

Albert Einstein

Contents

1 Introduction **1**
- 1.1 Shedding light on marine cyanobacteria . . . 1
- 1.2 Cyanophages are mixing up the oceanic gene pool . . . 5
  - 1.2.1 T7-like transcription system - a characteristic strategy of infection . . . 6
- 1.3 Architecture and regulation of a transcription unit . . . 9
- 1.4 Phylogenetic footprinting . . . 10
- 1.5 Transcriptional regulatory networks . . . 11
- 1.6 The circadian clock of cyanobacteria is unique . . . 13
  - 1.6.1 Roles of group 2 σ factors in circadian regulation . . . 15
- 1.7 The new era of small RNA molecules - tiny but mighty regulators . . . 16
  - 1.7.1 How to find them? . . . 18
  - 1.7.2 Are the newly identified ncRNAs functional? . . . 19
  - 1.7.3 Riboswitches: mRNA-enslaved ncRNAs . . . 19
  - 1.7.4 Are there regulatory RNAs in cyanobacteria? . . . 20
- 1.8 Aims of this work . . . 21

2 Materials and Methods **23**
- 2.1 Experimental Part . . . 23
  - 2.1.1 Cultures, cultivation, plasmids . . . 23
  - 2.1.2 DNA and RNA oligos . . . 26
  - 2.1.3 Genetic manipulation of marine *Synechococcus* . . . 26
  - 2.1.4 Isolation of nucleic acids . . . 27
  - 2.1.5 DNA gel electrophoresis, Restriction, Ligation . . . 29
  - 2.1.6 RNA gel electrophoresis, Northern blotting and Hybridisation . . . 29
  - 2.1.7 Polymerase chain reaction (PCR) . . . 32
  - 2.1.8 Sequencing of DNA . . . 33
  - 2.1.9 Rapid amplification of cDNA ends (RACE) . . . 33
  - 2.1.10 Quantitative RT-PCR . . . 35
- 2.2 Bioinformatics part . . . 37
  - 2.2.1 Genome data . . . 37
  - 2.2.2 Software tools . . . 37
  - 2.2.3 Promoter prediction . . . 38
  - 2.2.4 Phylogenetic footprinting . . . 38
  - 2.2.5 Prediction of ncRNAs . . . 39

3 Results **42**
- 3.1 Cyanobacterial promoters . . . 42
  - 3.1.1 *psbD* promoter region . . . 43
  - 3.1.2 *csoS1-rbcLS* promoter region . . . 44
  - 3.1.3 *atpB* promoter region . . . 45
  - 3.1.4 *sfsA* downstream region . . . 45

3.2 Phylogenetic footprinting . . . 47
3.2.1 Orthologous gene sets . . . 48
3.2.2 Predicted TF binding sites and regulons . . . 48
3.2.3 Experimental verification of transcription initiation next to putative binding sites . . . 52
3.3 Circadian rhythm and *kai* genes . . . 53
3.4 Transcriptional regulation of the cyanophage P-SSP7 . . . 57
3.4.1 Mapping of 5' ends by RACE experiments . . . 59
3.4.2 Putative motif upstream of genes 1, 3 and 29 . . . 60
3.4.3 Transcription of phage-specific RNA polymerase . . . 62
3.4.4 Missing promoter site upstream gene 30 . . . 63
3.5 Small RNAs in marine and other cyanobacteria . . . 63
3.5.1 Known and abundant RNAs of marine strains . . . 63
3.5.2 Computational screening identified novel RNA species . . . 64
3.5.3 Experimental verification of predicted ncRNAs in *Prochlorococcus* Med4 . . . 65
3.5.4 Yfr1 exists in diverse cyanobacteria . . . 69
3.5.5 Conservation of Yfr7/6Sa across the whole cyanobacterial lineage . . . 70
3.5.6 Excursus to *Synechocystis* PCC 6803: A *cis*-encoded antisense RNA for *isiA* 78

4 Discussion **81**
4.1 Promoter architecture and transcriptional regulation . . . 81
4.1.1 The conserved regulatory potential of four marine cyanobacteria . . . 82
4.1.2 Circadian rhythm: clock proteins and other periodosome components . . . 85
4.1.3 Transcriptional signals of cyanophage P-SSP7 infecting *Prochlorococcus* Med4 86
4.2 New families of small RNAs in cyanobacteria . . . 87
4.2.1 Computational screening in marine genomes . . . 88
4.2.2 Non-coding RNAs of *Prochlorococcus* Med4 and their orthologs . . . 89
4.2.3 Ubiquitous presence of 6S RNA in cyanobacteria and clues about its function 90
4.2.4 A novel antisense RNA to *isiA* mRNA in *Synechocystis* PCC 6803 . . . 91
4.3 Conclusion and outlook . . . 93

Bibliography **95**

A Appendix **113**

List of Figures

1.1 Transmission electron micrograph of an ultrathin section of *P. marinus* SS120 (Bryant, 2003). 1
1.2 Depiction of the amount of chlorophyll present in the oceans and the amount of vegetation on land (Bryant, 2003). . . . 2
1.3 False color image of sea surface temperature during the Atlantic Ocean meridional transect cruise and cross sections along the transect (Johnson et al., 2006). . . . 3
1.4 Relationships between *Prochlorococcus* and other cyanobacteria inferred using 16S rDNA (Rocap et al., 2003) . . . 4
1.5 Genome arrangement of *Prochlorococcus* podovirus P-SSP7 (Sullivan et al., 2005) . . . 7
1.6 Synthesis and processing of T7 messenger RNAs (Dunn and Studier, 1983) . . . 8
1.7 Overview of the transcriptional regulatory network in *E. coli* (Martinez-Antonio and Collado-Vides, 2003). . . . 12
1.8 The cyanobacterial periodosome model (Bell-Pedersen et al., 2005). . . . 14
1.9 A model for *fhlA*-OxyS interaction (Argaman and Altuvia, 2000). . . . 17
1.10 Schematic representation of the proposed mechanism for TPP-dependent deactivation of *thiM* translation (Winkler et al., 2002). . . . 20

2.1 Cultures of *Prochlorococcus* and *Synechococcus*. . . . 24
2.2 PCR result of 5' RACE for *psbD* in *Prochlorococcus* SS120. . . . 34
2.3 Pipeline for comparative prediction of non-coding RNAs (Axmann et al., 2005). . . . 41

3.1 Raster-score-filter method (Vogel et al., 2003a) and weblogo of *Prochlorococcus* SS120 -10 boxes. . . . 42
3.2 Arrangement of *ppiA, ycf4, psbD* and *psbC* and alignment of upstream regions of *psbD* for five cyanobacterial strains. . . . 44
3.3 Comparison of the *csoS1-rbcLS* region from marine cyanobacteria. . . . 45
3.4 Arrangement of genes around *atpB* and alignment of *atpB* upstream sequences for five marine cyanobacteria. . . . 46
3.5 Synteny of *mviN*, *sfsA*, *amt1* and *lytB* and alignment of *sfsA* downstream sequences for five marine strains. . . . 46
3.6 Overview of orthologous protein-coding genes including σ and regulatory factors of four marine genomes. . . . 47
3.7 Results of the PCR step during 5' RACE experiments for *lexA, umuD*, PMM1427, *urtA, glnA, ntcA* in *Prochlorococcus* Med4. . . . 52
3.8 Arrangement of *rpl27* and *rpl21*, encoding 50S ribosomal proteins, and *kaiA, kaiB* and *kaiC* encoding the core clock proteins of cyanobacteria. . . . 54
3.9 Alignment of coding regions of *kaiA* from *Synechococcus* PCC 7942, WH 8102 and 7803 together with the spacer regions between *kaiB* and *rpl21* of *Prochlorococcus* strains, MIT 9313 and 9303. . . . 55
3.10 Alignment of the conserved *rpl21* upstream regions of *Synechococcus* and *Prochlorococcus*. . . . 56
3.11 5' RACE experiments for the cyanophage P-SSP7. . . . 59
3.12 Summary of identified sites and sequence comparisons for P-SSP7. . . . 61
3.13 Small RNAs in marine Cyanobacteria (Axmann et al., 2005). . . . 64
3.14 Experimental screen for the presence of an RNA-coding gene in the *guaB-trxA* intergenic region. . . . 67

3.15 Determination of half lifes for Yfr 1, Yfr2 and Yfr5 to 7 in *Prochlorococcus* Med4 (Axmann et al., 2005). 68

3.16 Test of transcript accumulation of Yfr1-7 from Med4 (MED) under different conditions (Axmann et al., 2005). 69

3.17 Comparison of Yfr2, Yfr3, Yfr4 and Yfr5 from Med4. 70

3.18 Identification of Yfr1 RNA in diverse cyanobacteria. 71

3.19 Identification of Yfr7 RNA in different marine cyanobacteria. 72

3.20 Identification of 6Sa RNA in different cyanobacteria. 73

3.21 6Sa/Yfr7 Parsimony tree based on the alignment of 13 sequences from diverse cyanobacteria and genome location of *ssaA* in these strains. 74

3.22 Identification and transcript accumulation of Yfr7 in *Prochlorococcus* Med4. 75

3.23 Knock-out construct for Yfr7 in *Synechococcus* WH 8102 and analysis of mutants. 76

3.24 Accumulation of different transcripts from the *isiAB* region under iron limitation and during high-light conditions. 78

3.25 Characterisation of an antisense RNA in *Synechocystis* PCC 6803. 79

4.1 Influence of IsrR on the accumulation of trimeric PSI and PSI-IsiA supercomplexes (Duehring et al., 2006). 93

2 Sequence logos of the best conserved elements obtained by phylogenetic footprinting analysis of *Prochlorococcus* Med4, MIT 9313, SS120 and *Synechococcus* WH 8102. 113

3 List of high scoring clusters revealed by RNA prediction in marine cyanobacteria (Axmann et al., 2005). 121

4 6Sa/Yfr7 alignment of 13 sequences from diverse cyanobacteria. 122

List of Tables

2.1 DNA oligos used for PNK labelling and hybridisation to Northern blots. 31

2.2 DNA oligos used for PNK labelling and hybridisation to Northern blots, which did not yield signals. 32

2.3 PCR primer list. 33

2.4 DNA oligos used for 5' RACE experiments analysing ncRNAs. 34

2.5 DNA oligos used for 5' RACE experiments for cyanophage P-SSP7. 35

2.6 DNA oligos used for 5' RACE experiments. 36

2.7 *Prochlorococcus* Med4 gene-specific RT-PCR primers. 37

3.1 Experimentally determined mRNA 5' ends of *Prochlorococcus* SS120. 43

3.2 Top five of predicted motifs and their best hits identified in *Prochlorococcus* Med4, SS120, MIT 9313 and *Synechococcus* WH 8102 via phylogenetic footprinting analysis. 49

3.3 Overview of clock gene orthologs in *Prochlorococcus* Med4, SS120, MIT 9313, *Synechococcus* WH 8102 in comparison to *Synechococcus* PCC 7942, the circadian clock model. 56

3.4 Summary of transcriptional and other signals mapped to the cyanophage P-SSP7 genome. . 58

3.5 Summary of identified ncRNA genes in *Prochlorococcus* MED4 and their orthologues in three related strains of marine cyanobacteria. 66

1 A core set of 35 regulatory factor and 5 σ factor genes conserved between *Prochlorococcus* Med4, MIT 9313, SS120 and *Synechococcus* WH 8102. 114

2 List of predictions for putative TF binding sites in marine cyanobacteria. 120

Chapter 1

Introduction

1.1 Shedding light on marine cyanobacteria

Today it seems, we know more about the outer space than about the oceans of our Earth covering nearly 75 % of its surface. Thus, it was not surprising that no more than 20 years ago, a new type of photosynthetic picoplankton was discovered in the open sea (Chisholm et al., 1988) which since has changed the understanding of marine food webs as well as of the global geochemical carbon cycle completely (Falkowski, 2002). Because of their small cell sizes with diameters of only 0.5 to 0.7 μm and low cell densities, these tiny cyanobacteria had not been discovered until sensitive shipboard flow cytometry became available. A distinctive feature of all cyanobacteria known at that time was the presence of a phycobilisome-type light-harvesting complex, for which, however, no evidence was found in these organisms. This fact, together with the unusual combination of pigments, and a typical prokaryotic ultrastructure, suggested that these picoplankters are free-living relatives of the prochlorophyte *Prochloron* (Chisholm et al., 1988), which performs oxygenic photosynthesis using chlorophyll b, like land plants and green algae (Chlorophyta) (Tomitani et al., 1999). Taking also the coccoid shape (Fig. 1.1) into account, the name of the new genus was decided: *Prochlorococcus* (Chisholm et al., 1992).

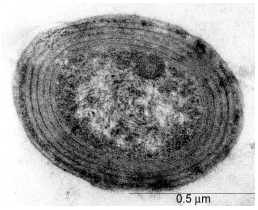


Figure 1.1: Transmission electron micrograph of an ultrathin section of *P. marinus* SS120, taken from Bryant (2003), showing the closely appressed thylakoids that distinguish this organism from marine *Synechococcus* sp. (Photograph courtesy of William Li and Frederic Partensky).

Fossil and molecular studies suggested that the archetype of cyanobacteria thrived successfully for billions of years (Schopf, 1993). Photosynthesis evolved early in these organisms, an activity that resulted in the enrichment of the planetary atmosphere in oxygen, so that

about 1.5 billion years ago our vital atmosphere had been created. As an endosymbiont the ancient cyanobacteria were transformed into an intracellular organelle - the chloroplast - without which no eukaryotic photoautotroph can exist, among them all higher plants (Douglas, 1998; Moreira et al., 2000). As a consequence of gene transfer from the endosymbiont to the original host's genome, about one fifth of all nuclear genes of extant land plants might descend from the cyanobacterial endosymbiont (Rujan and Martin, 2001; Martin et al., 2002). However, the known prochlorophytes, including *Prochlorococcus marinus*, have been shown to be not the specific ancestors of chloroplasts (Tomitani et al., 1999).
Today all life on Earth depends on photosynthesis. Now we are recognising that marine phytoplankton accounts for nearly 50 % of the net primary productivity of the biosphere (Field et al., 1998) which was grossly underestimated until recently. Satellite-based remote sensing (e.g., NASA sea-wide field sensor) has allowed more reliable determinations of oceanic photosynthetic productivity to be made (Field et al., 1998; Falkowski, 2002; Bryant, 2003); see Figure 1.2.

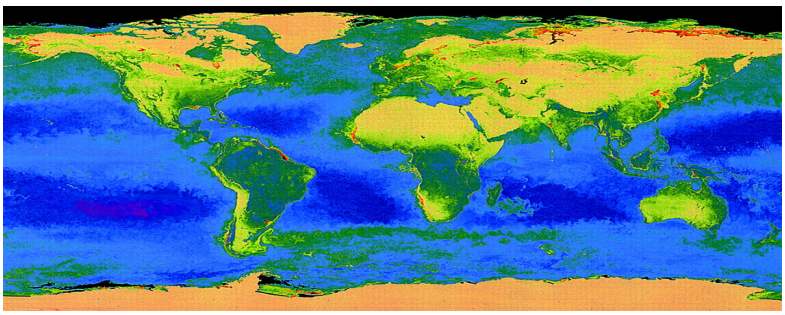


Figure 1.2: Depiction of the amount of chlorophyll present in the oceans and the amount of vegetation on land, taken from Bryant (2003). Purple and blue represent low levels of chlorophyll, and green, yellow, and red indicate progressively higher concentrations. Brown pixels show areas with little vegetation on land, and blue-green pixels represent areas of densest vegetation (Bryant, 2003). [Image provided by ORBIMAGE (Copyright 2003, Orbital Imaging Corporation) and processing by NASA Goddard Space Flight Center.]

Cyanobacteria today form a huge and heterogenous group of prokaryotes which is different in many features from other bacteria. They populate widely diverse environments such as freshwater, the oceans, hot springs, the surface of rocks, desert soil or the Antarctic. The group of *Prochlorococcus* and the closely related *Synechococcus* is abundant in the oceans and belongs to the most important primary producers as it could be responsible for nearly one-third of the primary biomass production on Earth (Fig. 1.2).
Total genome sequences are available for *Prochlorococcus marinus* SS120 (Dufresne et al., 2003), *Prochlorococcus* Med4 and *Prochlorococcus* MIT 9313 (Rocap et al., 2003). These strains are representative for distinct populations found in particular ecological niches within the marine ecosystem. The genus *Prochlorococcus* is often present at high abundances with more than 10^5 cells per ml in nutrient-poor areas of the world's oceans splitting up in two major ecotypes (Fig. 1.4) - one being represented by the high-light-adapted strains such as Med4, the other by low-light-adapted strains SS120 and MIT 9313 (Moore et al., 1998; Partensky et al., 1999; West and Scanlan, 1999; Fuhrman and Capone, 2001).

These sometimes co-occurring ecotypes have not only different light optima for growth, pigment contents and light-harvesting efficiencies but also different temperature optima (Johnson et al., 2006), sensitivities for trace metals, and are each specifically affected by cyanophages (Rocap et al., 2003; Moore et al., 1998; Sullivan et al., 2003). Nevertheless, on the basis of their rDNA similarity they would be recognised as a single species as their ribosomal DNA sequences differ by less than 3 % (Hagstrom et al., 2002).
Very recently, the variable *Prochlorococcus* ecotype distribution was correlated to environmental gradients along a meridional transect in the Atlantic Ocean, showing that temperature plays a dominant role in regulating some ecotype distributions (see Fig. 1.3). Other factors as light and competitor abundance (*Synechococcus*) can also be important, while nutrient availability seems to play a minor role (Johnson et al., 2006).

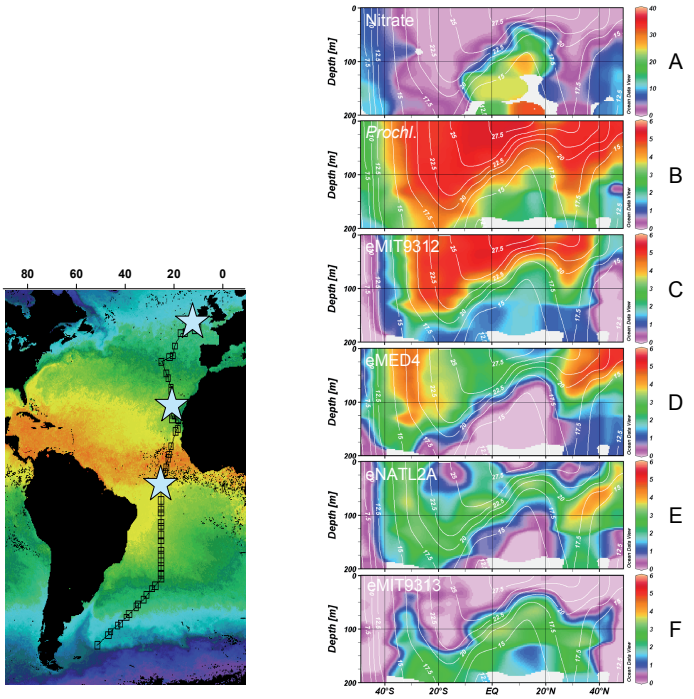


Figure 1.3: Left panel: False color image of sea surface temperature during the Atlantic Ocean meridional transect cruise (AMT13), represented by Johnson et al. (2006), during Sept/Oct 2003 overlaid with cruise track and sampled stations (boxes). Reds are warm temperatures, blues are cold temperatures. Right panel: Cross section along the transect (upper 200 m) of (A) nitrate concentration (μmol/kg), (B) total Prochlorococcus (log cells/ml) and four dominant ecotype clade abundances (log cells/ml): (C) eMIT9312, (D) eMED4, (E) eNATL2A, and (F) eMIT9313. Temperature contour lines are shown in white. The lower limit of detection is ∼1 cell/ml (Johnson et al., 2006).

The next sequenced relative to the group of *Prochlorococcus* (Fig. 1.4) is the strain *Synechococcus* WH 8102 (Urbach et al., 1992), which has evolved a unique type of swimming motility (Palenik et al., 2003). Marine *Synechococcus* species, first described in 1979 (Waterbury et al., 1979; Johnson and Sieburth, 1979), are typically less abundant in very oligotrophic environments, but have a broader global distribution usually in the upper

layer of the ocean, where white or blue-green light is available for photosynthesis (Ting et al., 2002; Bryant, 2003). Known so far, all marine *Synechococcus* strains harbour chlorophyll a (and no chlorophyll b) and phycobilisomes (Waterbury et al., 1986), but a great genetic diversity exits within this genus as well, with strains probably adapted to ecological niches, whereby WH 8102 is more of a generalist with adopted strategies for conserving limited iron stores and a reduced regulatory machinery (Palenik et al., 2003).
Compared to *Synechococcus* and other cyanobacteria, strains of *Prochlorococcus* are unusual in lacking phycobilisome light-harvesting complexes. Instead chlorophyll-protein complexes known as Pcbs are used, which contain divinyl chlorophyll a and chlorophyll b (La Roche et al., 1996; Bibby et al., 2001b; Hess et al., 2001).

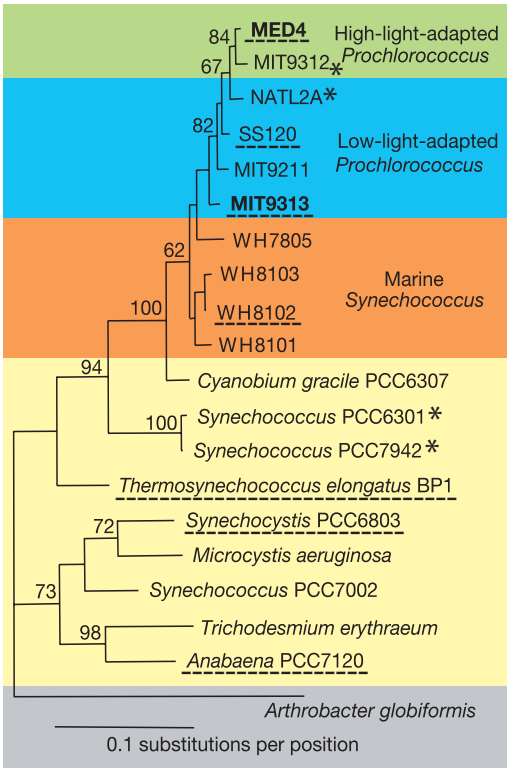


Figure 1.4: Relationships between *Prochlorococcus* and other cyanobacteria inferred using 16S rDNA, shown by Rocap et al. (2003). Strains with dashed lines are sequenced, annotated and published (Dufresne et al., 2003; Rocap et al., 2003; Palenik et al., 2003; Nakamura et al., 2002; Kaneko and Tabata, 1997; Kaneko et al., 2001). For strains labelled with a star complete genome sequence is available at NCBI database (Wheeler et al., 2005). Unfinished sequences by whole genome shotgun sequencing projects exist for several other strains at NCBI (Wheeler et al., 2005).

The information of at least four fully sequenced and annotated genomes of closely related and even so ecologically divergent marine cyanobacteria offers a great chance for comparative studies. The conserved sequence motifs and patterns essential for regulation as well as individual regulatory signals and specific gene contents responsible for niche differentiation could be filtered out and studied in detail. All sequenced strains are available for cell

culturing, and a genetic manipulation system has been developed for *Synechococcus* WH 8102 (Brahamsha, 1996). Thus, not only a computational analysis of these cyanobacterial genomes appears promising, but these can be directly linked to corresponding experiments in the laboratory to verify and improve the *in silico* predictions.
That could be a starting point to understand how these tiny and even so specialised organisms could dominate the oceans for millions of years although environmental conditions were and are changing. A major outcome of the total genome analyses of these cyanobacteria has been their small and compact genomes, being considered as the minimal genome of a free-living photoautotroph. But, which regulatory potential does remain in such a streamlined bacterium to enable responses to variations in temperature, light condition, nutrient or metal? Today the global warming is faster than it was assumed initially - to what extent the complex marine ecosystem will be affected? Can the regulatory and genetic potential of these microbes sustain adaptation to a big global change? At which point will it suddenly and unexpectedly crash, with a major impact for our biosphere?

1.2 Cyanophages are mixing up the oceanic gene pool

The analysis of four genomes of marine cyanobacteria has revealed an unprecedented degree of genomic variation. What processes could alter microbial genotypes and enable such extensive niche differentiation as observed within the group of slow-growing *Prochlorococcus* and *Synechococcus* in the ocean? The annotation of the sequenced genomes yielded INT family site-specific recombinases, suggesting that prophages were once integrated. The comparison and analysis of very closely related genomes as *Prochlorococcus* Med4 and MIT 9313 revealed numerous large and small-scale rearrangements, where the break points between the orthologous gene clusters are often flanked by tRNAs (Rocap et al., 2003). It was already suggested that tRNA genes (and tmRNA genes) are common integration sites for phages and mobile elements (Williams, 2002).
Indeed, cyanophages were isolated that infect *Prochlorococcus*. Thereby, *Podoviridae* normally infecting high-light-adapted *Prochlorococcus* were found to be extremely host-specific (Sullivan et al., 2003). For example, P-SSP7 exclusively infects the high-light-adapted *Prochlorococcus* Med4 strain. On the other hand, low-light-adapted *Prochlorococcus* as well as all *Synechococcus* strains yielded primarily *Myoviridae* (Sullivan et al., 2003).

Genome sequencing of several cyanophages uncovered another fascinating feature of cyanobacterial viruses: Besides established "core" genes known for each phage group, the cyanophage genomes contain not only a few host-homologous genes 'by accident', but even encode functional proteins, which are central to oxygenic photosynthesis and may be crucial to maintain photosynthetic activity during infection (Mann et al., 2003; Lindell et al., 2004; Millard et al., 2004). Some of the host genes that appear common in these phages are suggested to represent "signature" genes which could be specifically characteristic of oceanic cyanophages (Sullivan et al., 2005). For example, all marine cyanophage genomes (except podovirus P60) sequenced so far contain a *psbA* gene that encodes a core reaction center protein (D1) in photosystem II. Very recently it was shown that phage *psbA* and *hli* (high-light inducible) genes are expressed during infection of *Prochlorococcus* and that

they may increase phage fitness (Lindell et al., 2005).
Interestingly, the distribution of other phage-encoded signature genes is more sporadic among the two viral families (*Podoviridae* and *Myoviridae*) which may reflect adaptations for infection of photosynthetic hosts in low-nutrient oceanic environments (Sullivan et al., 2005). Thus, both T4-like myoviruses P-SSM2 and P-SSM4 harbour genes (*phoH, pstS*) that are likely to be important for responses to phosphate stress (Sullivan et al., 2005).

For bacteriophages an obligate lytic or temperate (lysogenic) lifestyle may be distinguished. Infection by a temperate phage may either cause lysis or entry of a stable state within the host (lysogeny) by integration of the phage genome into the host genome as a prophage, which could be induced back into the lytic cycle in response to environmental conditions (Mann, 2003).
Viral infection of marine unicellular cyanobacteria was first described in 1990 (Suttle et al., 1990; Proctor and Fuhrman, 1990), and up to now, temperate phages have not been identified in complete genome sequences of marine *Prochlorococcus* and *Synechococcus* strains, although some field studies dealing with lysogeny in natural populations suggest this is a possibility (Ortmann et al., 2002; McDaniel et al., 2002).
However, the phenomenon of pseudolysogeny was observed by studying interactions between phage and nutrient-starved *Synechococcus* host (Ripp and Miller, 1997; Wilson et al., 1996). Pseudolysogeny describes a phage-host relationship in which a phage-infected cell grows and divides even though its virus is pursuing a lytic infection (Birge, 2000).

The marine *Podoviridae* known so far have an obligate lytic lifestyle (lacking known lysogeny genes), although in the T7-like P-SSP7 genome a tyrosine site-specific recombinase (*int*) gene, which enables a temperate phage to integrate its genome into the host genome, as well as a 42-bp sequence identical to the 5' part of the leucine tRNA gene in the host genome (*Prochlorococcus* Med4) was discovered (Sullivan et al., 2005). The cyanophage P-SSP7 complete genome is available now (Sullivan et al., 2005) encoding 54 open reading frames (ORFs) within nearly 45 kb. It includes a T7-like RNA polymerase gene as well as the host-like *hli* and *psbA* genes (Fig. 1.5).

Half of the translated ORFs in P-SSP7 could not be assigned a function (Sullivan et al., 2005) which is a general problem for marine phage genomes. They have been described as the largest untapped reservoir of genetic information (Paul et al., 2005), and further sequences of marine phages as well as future studies are needed to understand the evolutionary history and biological and ecological functions of the marine viral community.

1.2.1 T7-like transcription system - a characteristic strategy of infection

The cyanophage P-SSP7 genome is most similar to genomes of the T7-like phages (Sullivan et al., 2005). The lytic phage T7 infecting *E. coli* cells is one of the best studied podoviruses. Members of this group (e.g. T7, T3, SP6) possess a unique strategy of infection, where the phage's own transcription system plays an important role in phage gene expression and also drives the efficient translocation of viral DNA into the host cells (Chen and Schneider, 2005). The T7 genome contains bacterial (host-like, class I) promoters as well as 17 promoters only recognised by their T7 RNA polymerase (RNAP) (Dunn and

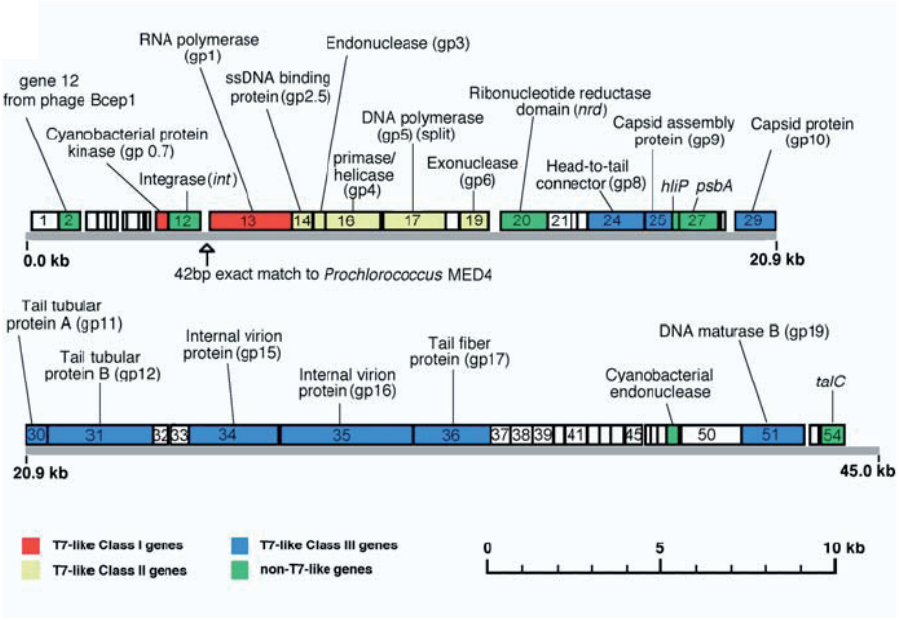


Figure 1.5: Genome arrangement of *Prochlorococcus* podovirus P-SSP7, described by Sullivan et al. (2005). The ORFs are sequentially numbered within the boxes, and gene names are designated above the boxes. Gene designations use T7 nomenclature for T7-like genes (Molineux, 2001) or microbial nomenclature for non-phage genes. Class I, II, and III genes refer to those in T7 that belong to gene regions primarily involved in host transcription of phage genes (class I), DNA replication (class II), and the formation of the virion structure (class III). The ORFs are designated by boxes, and in this genome, all ORFs are oriented in the same direction. Although the phage genome is one molecule of DNA, the representation is broken to fit on a single page. Note that the P-SSP7 genome is most similar to genomes of the T7-like phages (Sullivan et al., 2005).

Studier, 1983). The promoters for T7 RNAP were classified into three groups: class II, class III or replication promoters depending on their location, temporal utilisation and function during infection. Thus, T7 genes are transcribed from left to right and expressed co-ordinately in three groups as illustrated in Figure 1.6a (Dunn and Studier, 1983):
(I) At first, the host RNAP initiates transcription at strong bacterial promoters located near the left end of the bacteriophage DNA and transcribes the genome until the early termination site (TE), resulting in a segment of about 20 % of the phage genome, which includes the phage RNAP gene.
(II) During the next stage of infection, the newly-made phage RNAP specifically recognises its promoters and transcribes about 2/3 of the phage genome until a strong termination site (Tϕ) without any accessory factors.
(III) Termination at Tϕ is not completely efficient and would be lethal to T7 as there is no promoter for T7 RNAP between this termination site and the following genes, which specify structural proteins for T7, so these genes must be transcribed entirely by read-through of Tϕ. Thus, the class III genes are the last to be expressed.

Additionally, cleavage of T7 RNAs at specific sites is a prominent feature (Dunn and Studier, 1973a,b, 1975): Ten RNase III cleavage sites were located inside the T7 genome,

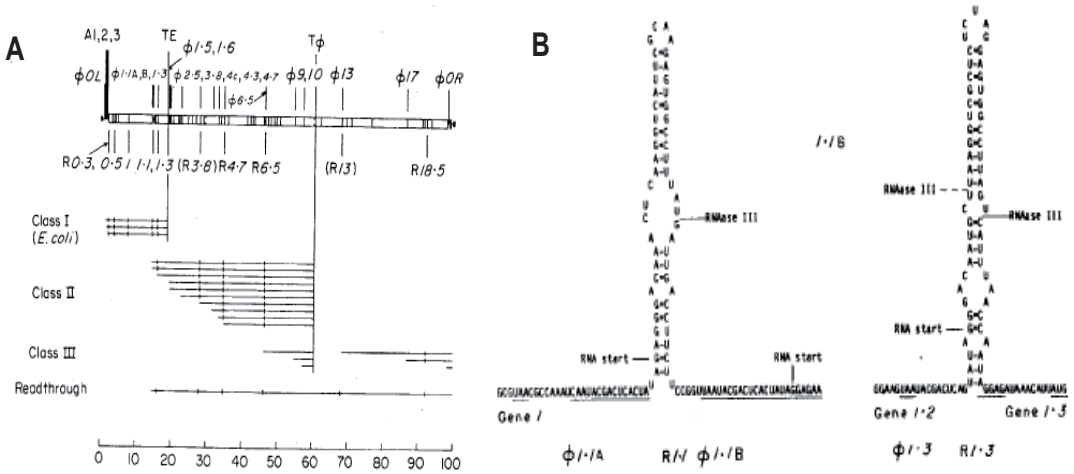


Figure 1.6: Synthesis and processing of T7 messenger RNAs summarised by Dunn and Studier (1983). (A) The T7 genes are represented by open boxes; the positions of transcription signals are given above the genes, the RNase III cleavage sites below. The primary transcript from each promoter is represented by a horizontal line, and sites of RNase III cleavage by short vertical lines. Apparently not all RNAs are cut at the R3.8 and R13 RNase III cleavage sites, as indicated by parentheses. RNAs produced by ready-through of Tϕ are also represented. The scale bar below is in map units. (B) Two examples for RNase III cleavage sites in T7 RNAs. Potential base-paired structures are indicated, as are the locations of the termination codons for the genes immediately preceding, and the ribosome-binding sequences and initiation codons for the genes immediately beyond the cleavage sites. The locations of the promoters for T7 RNAP are also shown. The known positions of RNase III cleavage for the two (of five) cleavages in T7 early RNA are indicated. The position of the secondary cleavage in R1.3 is indicated by a broken line (Dunn and Studier, 1983).

five in the early (Kramer et al., 1974; Rosenberg et al., 1974; Kramer and Rosenberg, 1976; Rosenberg and Kramer, 1977; Robertson et al., 1977) and five in the late region (Dunn and Studier, 1983). At these sites, the primary transcripts are processed by a host enzyme, RNase III, to provide the mRNAs observed *in vivo* (Dunn and Studier, 1983). Ribonuclease III from *E. coli* participates in maturation and decay pathways by site-specifically cleaving double-helical structures in cellular and viral RNAs, whereby the involvement of base-pair sequence in determining cleavage sites is unclear (Zhang and Nicholson, 1997). In the T7 genome, however, the promoter regions for T7 RNAP often overlap with sequences needed for the RNase III cleavage forming a characteristic pattern of base-pairing (Fig. 1.6b).

The expression of T7 RNAP early during infection and the inactivation of the host polymerase, efficiently directs all gene expression in the cell to T7 DNA (Zhang and Studier, 2004). T7 RNAP appears to have multiple roles in selective transcription of T7 genes as well as in replication, maturation and packaging (Zhang and Studier, 2004). During the last stage of infection, T7 lysozyme binds T7 RNAP and inhibits transcription, and stimulates replication and packaging of T7 DNA (Zhang and Studier, 1997, 2004). At the end, T7 lysozyme cuts a bond in the peptidoglycan layer of the cell wall to release

newly-made phages.

For the circular marine P-SSP7 genome (Fig. 1.5) a similar architecture compared to coliphage T7 was observed including the essential T7-like RNAP (Sullivan et al., 2005), which may implicate a similar infection strategy as described above. On the other hand, in a recent study about T7-like promoters and their RNAPs, no significant promoter sites were found for P-SSP7 suggesting that this phage does not belong to the highly related T7 group (including T7, T3, SP6) but rather to a more distantly related T7 supergroup (Chen and Schneider, 2005). Chen and Schneider (2005) proposed, that P-SSP7 may utilise a different transcriptional strategy to that known for T7, because P-SSP7 does not contain both components of the T7-like transcription system (RNA polymerase and its specific promoters). Further studies are needed focusing on P-SSP7 DNA transcription during Med4 lytic infection to identify different stages of gene expression as well as their responsible promoter sites.

1.3 Architecture and regulation of a transcription unit

A transcription unit is defined as the DNA between a promoter and a terminator that serves as a template for a complementary strand of RNA which can be messenger RNA (mRNA) encoding a protein, ribosomal RNA (rRNA), transfer RNA (tRNA) or small non-coding RNA (ncRNA) with regulatory functions. The bacterial RNA polymerase holoenzyme that synthesises these RNAs consists of two components - a core enzyme responsible for elongating the RNA chain and a σ factor required at the stage of initiation for promoter recognition. Bacteria contain several σ factors, a strategy that is in itself the first level of regulation of transcription initiation. Alternative σ factors activated by adverse conditions (heat shock, nitrogen starvation) can direct the core enzyme to recognise promoters with different consensus sequences. As new σ factors become active, old σ factors are displaced, that turns genes off and dictates when its set of target genes is expressed, and the amount of factor available influences the level of gene expression (Lewin, 2000; Rojo, 1999). More than one σ factor may be active at any time, and the specificities of some of σ factors overlap equipping the bacterial cell with a flexible regulatory network.

Transcription has been studied the best in *Escherichia coli* which uses seven different species of σ factors to modulate promoter activation. Based on the abundant σ factor of housekeeping genes (σ^{70} reviewed by Ishihama (2000)) a consensus sequence of σ^{70}-dependent promoters has been established, and is defined by two consensus hexamers, TTGACA and TATAAT, centred -35 and -10 base-pairs respectively upstream of the transcription initiation site (Hawley and McClure, 1983).

While the basic principles of promoter recognition and geometry of RNA polymerase seems to be similar for all eubacteria, the situation in cyanobacteria could be different by the presence of a truncated RNA polymerase subunit β' (γ) encoded by *rpoC1* (Xie et al., 1989) and an additional subunit β" (δ) encoded by *rpoC2*. These subunits correspond to the N- and C- terminal parts of the β' subunit of other eubacteria plus an additional protein domain of 70 kDa, and are found in this split form in all cyanobacteria including

Prochlorococcus as well as in plant plastids. Although RNA polymerases from enterobacteria and cyanobacteria share similar domains and subunits, it was shown that they are not directly interchangeable (Schyns et al., 1998). It would not be the first case, where known facts about *E. coli* cannot be extrapolated directly to any other bacterium, in this case to a cyanobacterium. On the other hand, only a few studies existed about factors required for the transcription of a gene in a cyanobacterium.
The first systematical study of transcription initiation sites was done in a marine cyanobacterium, *Prochlorococcus* Med4 where a subset of promoters was analysed to suggest a consensus promoter sequence and to predict additional promoters in the whole genome of Med4 (Vogel et al., 2003a). The computational analysis revealed some general features of the cyanobacterial promoter architecture. The alignment of the experimentally identified regions suggested a consensus promoter structure similar to other eubacteria whereby -10 region and TIS are spaced by about six base-pairs; transcription and translation start site are spaced by 15 to 85 base-pairs in most cases. The -10 box itself exhibits three conserved nucleotides: T(-12), A(-11) and T(-7) which partly represents the well-known consensus TATAAT of *E. coli* and *B. subtilis*. The -35 regions do not possess overall conservations but were found to define different subsets (Vogel et al., 2003a). One subset of promoters with a -35 consensus similar to the known TTGACA sequence for house-keeping genes of *E. coli* incloses the Med4 genes *kaiB, rpl21, rps12* and *ccmK* (Vogel et al., 2003a).
This knowledge leads us to additional regulatory sites beside the core promoters such as transcription factor binding sites (TFBS) and patterns responsible for a secondary structure of the 5' untranslated region (5' UTR) as in the case of recently found riboswitches.

1.4 Phylogenetic footprinting

Bacterial regulatory sites rarely occur as sequence patterns conserved over a certain length, but rather as spaced, short words, what makes them more tricky to detect in the matrix of millions of only four different letters A, T, G and C. Phylogenetic footprinting is the major method for enriching for candidate regulatory elements (Bulyk, 2003). This technique searches for conserved motifs upstream of orthologous genes from closely related species - a search for "islands of conserved sequences in seas of less conserved non-coding sequences" (Pennacchio and Rubin, 2001). Sequence similarity is again the foundation for this bioinformatic method assuming that mutations within functional regions of genes accumulate more slowly than mutations in regions without sequence-specific function (Wasserman and Sandelin, 2004). The existing phylogenetic footprinting algorithms can be divided in three components:

- defining suitable orthologous gene sequences for comparison

- aligning the promoter sequences of orthologous genes

- visualising or identifying segments of significant conservation

The great power of phylogenetic footprinting algorithms has been shown for organisms of all kingdoms of life as the prediction of transcription regulatory sites in Archaea by Gelfand et al. (2000), in 17 complete microbial genomes of *E. coli* and related γ-proteobacteria by McGuire et al. (2000) and others, recently for the metal reducing bacterial family *Geobacteraceae* (Yan et al., 2004), as well as for yeast (Cliften et al., 2001), mouse and human (Loots et al., 2000; Gottgens et al., 2001), for example. Reviews of methods and available resources are given by numerous articles (Bulyk, 2003; Wasserman and Sandelin, 2004; Ureta-Vidal et al., 2003; Frazer et al., 2003; Dubchak and Frazer, 2003). Thereby, the initial and maybe the most challenging step is selecting a set of genomes with the appropriate level of nucleotide similarity (evolutionary distance) of the sequences. One has to balance the need to align orthologous sequences with the aim of having the functional elements standing out (Cliften et al., 2001). The genomes of the four closely related but differentially adapted marine cyanobacteria, *Prochlorococcus* Med4, SS120, MIT 9313 and *Synechococcus* WH 8102, may represent the right evolutionary distance to implement a phylogenetic footprinting approach. Their genomes come close to the minimal genome size for free-living autotrophs, so their regulatory machinery has been under intensive streamlining pressure. Thus, a comparative analysis could provide clues about the minimal regulatory equipment of a marine picocyanobacterium mediated by already known and hitherto unidentified genes and regulatory sequences like DNA binding sites of transcriptional factors.

1.5 Transcriptional regulatory networks

Cells have to make decisions, when environmental conditions are changing. Their decision-making process includes protein-DNA interactions defined by transcriptional factors (TF) and their targets around promoters. How does the bacterial cell sense the changes in environmental conditions which may start a signal-transduction pathway? Transcription factors are mainly two-domain proteins, consisting of a DNA-binding domain along with a regulatory domain (Madan Babu and Teichmann, 2003). The regulatory domains are often sensors for small metabolites mediating precise activation or repression of the genes belonging to the regulon of the TF. Other TFs are part of a two-component signal transduction system where a CheY-like response regulator receiver domain is phosphorylated by kinases.

The best known transcriptional network for any cell is the one of *E. coli* where about 300 genes (8 %) of 4405 identified ORFs are predicted or known TFs (Blattner et al., 1997; Perez-Rueda and Collado-Vides, 2000). An overview of the complex network of currently known interactions is given in Figure 1.7 where only seven regulatory proteins (CRP, FNR, IHF, FIS, ArcA, NarL, and Lrp) directly modulate the expression of more than the half of all genes in *E. coli*. On the other extreme, about one fifth of all transcription factors in *E. coli* regulate only one or two genes (Martinez-Antonio and Collado-Vides, 2003). The pleiotropic phenotype and the ability to regulate operons that belong to different pathways distinguish global regulators from local or dedicated regulators (Gottesman, 1984). However, in many cases regulation occurs by multiple TFs and often a global regulator acts together with more specific local regulators. The fact that three-quarters of TFs have

arisen by gene duplication (Madan Babu and Teichmann, 2003) gives an idea how these complex regulatory networks had been developed.

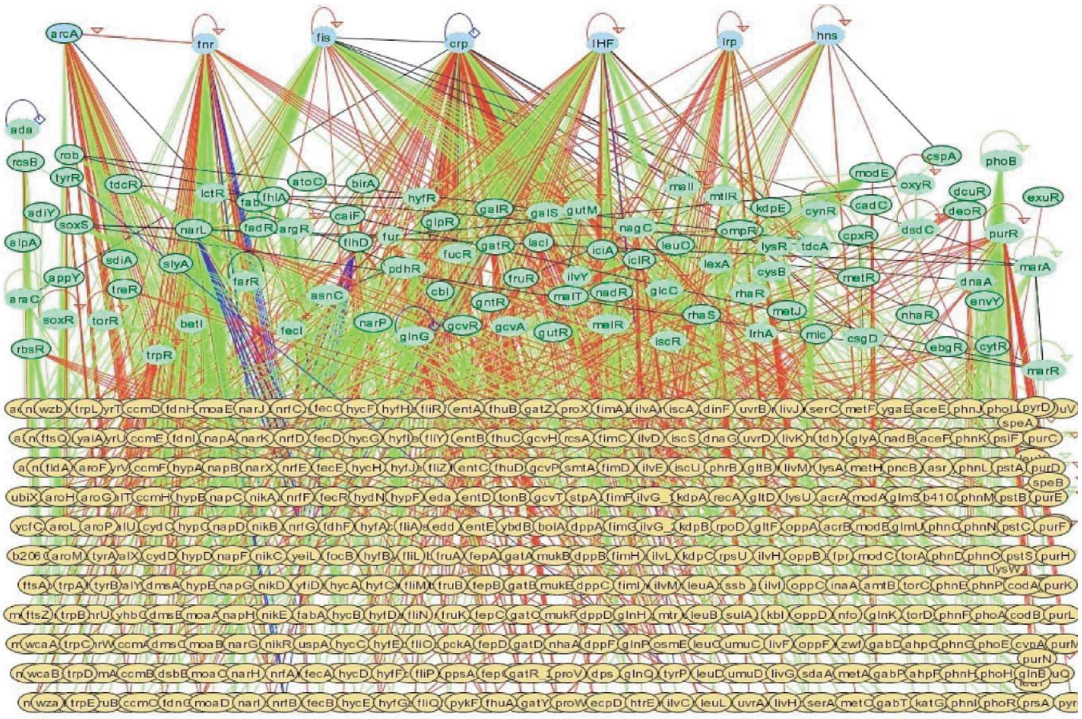


Figure 1.7: Overview of the transcriptional regulatory network in *E. coli* suggested by Martinez-Antonio and Collado-Vides (2003). Regulated genes are shown as yellow ovals, TFs are shown as green ovals and TFs considered to be global regulators are shown as blue ovals. The green lines indicate activation, red lines indicate repression and dark blue lines indicate dual regulation (activation and repression) (Martinez-Antonio and Collado-Vides, 2003).

In a regulatory network not only the number of interactions has to be considered but also a hierarchy of different levels of regulation. In a bacterial cell where TFs regulate TFs, including themselves in feedback loops (auto-regulation is frequent) and complex relationships, a hierarchical organisation of the network can not be assumed from these figures (Martinez-Antonio and Collado-Vides, 2003). In general, regulators work together with other regulators in a given condition resulting in groups of co-regulated genes, which provides the cell the flexibility to tune its transitions. Additionally, a single regulator can be capable of regulating genes transcribed by different σ factors and σ factors regulate genes subject to specific TFs. Summing all up, the cell is able to partition its genes by alternative ways, either by relationships between different TFs, or in terms of classes of genes of different σ factor promoters (Martinez-Antonio and Collado-Vides, 2003).

Although the bacterial regulation network alone is a very complex network, it can be broken down into basic building blocks and 'network motifs' can be defined (Shen-Orr et al., 2002), revealing that there is only a small number of multi-level regulatory cascades

amongst the transcription factors (Madan Babu and Teichmann, 2003). A large part of the network is composed of repeated appearances of three highly significant network motifs. This and related approaches help to define and to implement the basic computational elements of biological networks generating models which may be supportive in gaining insights into dynamic behaviours (Shen-Orr et al., 2002).

1.6 The circadian clock of cyanobacteria is unique

Daily oscillation mediated by an internal clock are one of the most fascinating biological non-linear dynamics found within a single cell. Cyanobacteria are the simplest organisms known to possess a true circadian clock, and in particular *Synechococcus elongatus* is used as an instructive model system for circadian mechanisms in unicellular organisms (Iwasaki and Kondo, 2004). The temporal coordination of internal biological processes and adaptation to daily fluctuations in the environment play a critical role in survival of diverse organisms, from bacteria to humans (Bell-Pedersen et al., 2005; Kucho et al., 2005). Circadian clocks exist in all eukaryotes (except yeast) and were longtime thought to be restricted to the eukaryotic kingdom. Today their existence has been demonstrated for plants, animals and, among prokaryota, exclusively in cyanobacteria. The design principles are conserved between these highly different groups of organisms, whereas their components share no similarity. At the core of all circadian clocks is at least one internal autonomous circadian oscillator containing positive and negative elements that form autoregulatory feedback loops (Young and Kay, 2001).

Defining properties of circadian rhythms include (Bell-Pedersen et al., 2005):

- a periodicity of about 24 hours, even in the absence of an environmental cycle (free-running rhythm)

- the ability of the clock to be entrained in a time-dependent manner by environmental stimuli

- the compensation of periodic length for changes in the natural environment, in particular temperature compensation

Studies regarding the outright incompatibility of cyanobacterial nitrogen fixation and oxygenic photosynthesis showed that temporal separation of disparate processes is advantageous for the cell and suggested an endogenous timing mechanism for cyanobacteria (Golden et al., 1997). An authentic circadian clock has been confirmed in *Synechococcus elongatus* PCC 7942 which fulfils all three criteria of a circadian process described above; although PCC 7942 does not fix nitrogen.

In *Synechococcus elongatus* a pacemaker orchestrates a global rhythmic regulation of gene expression and controls the timing of cell division (Golden and Canales, 2003; Johnson,

2004), but the circadian timing-mechanism itself is independent of the cell cycle (Kondo et al., 1997). The cyanobacterial pacemaker is based on the clock proteins KaiA, KaiB and KaiC which are together with ATP sufficient to achieve a temperature-compensated circadian rhythm of KaiC phosphorylation *in vitro* (Nakajima et al., 2005). The fundamental timekeeping mechanism involves interactions among clock proteins rather than transcriptional control, culminating in the formation of a high-molecular-weight complex during the night termed as the periodosome (Bell-Pedersen et al., 2005); see also Figure 1.8.

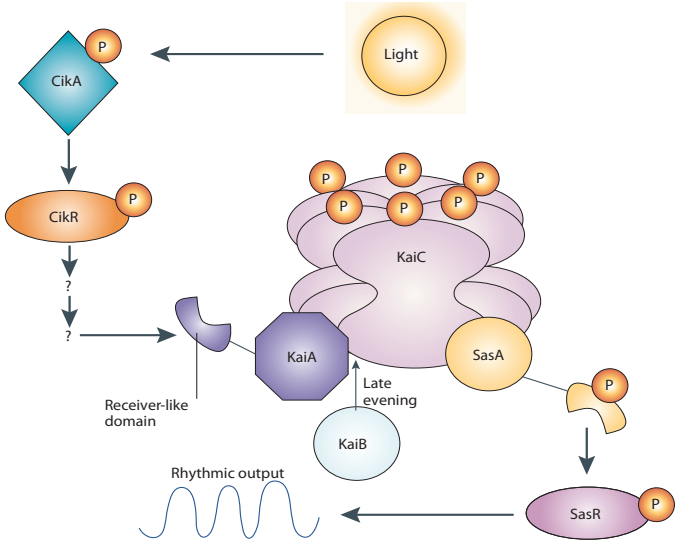


Figure 1.8: The cyanobacterial periodosome model, reviewed by Bell-Pedersen et al. (2005). Environmental information, such as daylight, is transduced through the phosphorylation and activation of Circadian input kinase A (CikA). CikA in turn phophorylates and activates its predicted binding partner, Circadian input kinase R (CikR). Information is then transferred through protein-protein interactions to the receiver-like domain of the circadian-clock protein KaiA. KaiA interacts with KaiC and stimulates autophosphorylation of KaiC, which is hexameric. In the phosphorylated state, KaiC hexamers can form a complex with other clock components. *Synechococcus* adaptive sensor A (SasA) joins the complex and is thereby stimulated to phosphorylate its predicted binding partner, *Synechococcus* adaptive sensor R (SasR). Phosphorylated, active SasR sends temporal information from the periodosome to the rest of the cell to activate rhythmic gene expression, either directly or indirectly. Late in the evening, another protein, KaiB, joins the periodosome introducing its disassembly and restarting the cycle. The modular events that reactivate the cycle in constant environmental conditions have not been yet described (Bell-Pedersen et al., 2005).

In a new model for the core clockwork in cyanobacteria rhythmic changes of chromosome condensation underlie the rhythms of gene expression (Johnson, 2004). Recently, it was suggested that the periodosome interacts directly with the bacterial nucleoid to affect supercoiling (Iwasaki and Kondo, 2004; Xu et al., 2003), but an experimental verification is still missing. However, the supercoiling level in *Chlamydomonas* chloroplasts, evolutionary a plant organelle derived from cyanobacteria, was found to be highest at the beginning and lowest at the end of the light phase, indicating an increase in DNA supercoiling activity

at night (Salvador et al., 1998). Circadian changes of the chromosome topology in the *Synechococcus* remain to be demonstrated.

Moreover, an intimate link to the metabolic status of the cell seems to exist. For example, the light-dependent protein A contains redox-active iron-sulfur clusters and is, therefore, sensitive to the redox state of the cell (Ivleva et al., 2005). LpdA co-purifies with periodosome components and forms a complex with clock proteins in a circadian fashion (Ivleva et al., 2005). It was shown that LdpA affects the absolute, as well as the light- and redox-dependent, abundance of CikA, a key input pathway component, and furthermore the absolute level of the clock protein KaiA. Thus, a novel input pathway to the circadian oscillator is suggested in which LdpA is a component of the clock protein complex that senses a metabolic state of a cell for adjusting the period length according to light intensities, but also to other environmental factors, such as nutrients and temperature (Ivleva et al., 2005).

Today, it is known that the cyanobacterial clock controls many physiological processes, such as photosynthesis, amino acid uptake, carbohydrate synthesis, and the cell division cycle (Golden et al., 1997; Iwasaki and Kondo, 2004). Although the entire genome seems to be under clock control at the transcriptional level, different sets of genes are expressed with distinct phase relationships, indicating that there are also other layers of control (Bell-Pedersen et al., 2005). In some mutant backgrounds different periods can be observed for different genes. Together, all the available data indicate that multiple circuits coexist in the cell and that more than one oscillator is likely to function within the cyanobacterial cell (Bell-Pedersen et al., 2005; Iwasaki and Kondo, 2004).

1.6.1 Roles of group 2 σ factors in circadian regulation

Cyanobacteria are unusual in having multiple, closely related, σ factors that are not essential for cell viability (group 2) in addition to the essential, principal σ factor, RpoD1 (Imamura et al., 2003a). A model was suggested in which σ factors work as a consortium to convey circadian information to downstream genes: Because each σ factor dissimilarly affects transcription from specific cyanobacterial promoters, the cyanobacterial transcriptional apparatus may oscillate in a circadian manner (oscillations in RpoD4 protein level have been demonstrated) by changing the composition of the RNA polymerase holoenzyme for individual group 2 σ factors over the circadian cycle (Nair et al., 2002; Ditty et al., 2003).

Inactivation of any of the four known group 2 σ factor genes (*rpoD2, rpoD3, rpoD4* and *sigC*), either single or in pairs, alters circadian expression of the *psbA* promoter in *Synechococcus* PCC 7942 (Nair et al., 2002). Two most striking examples are mutants lacking one of the group 2 σ factors, SigC or RpoD3, which results in lengthening or shortening the period of the gene expression rhythm of *psbA* without affecting the period of *kaiBC* expression (Nair et al., 2002). These results reveal the possibility that multiple timing circuits are coexisting at least in *Synechococcus elongatus* (Nair et al., 2002).

The group 2 σ factor RpoD3 of *Synechococcus* PCC 7942 was assigned to SigD in *Synechocystis* PCC 6803 (Imamura et al., 2003b), which is light-induced and responsible for

promoter recognition of the photosynthesis gene, *psbA*. Studies by Imamura et al. (2003a,b, 2004) in *Synechocystis* PCC 6803 revealed an antagonistic expression of the group 2 σ factors SigB and SigD. These are inversely regulated with respect to light/dark changes and depend on the redox status of the photosynthetic electron transfer chain. Moreover, in *sigB* and *sigD* knockout strains, the expression was significantly reduced of the dark-induced *lrtA* as well as the light-induced *psbA2/3* transcript, respectively. A possible model for these observations includes repression of SigB expression by SigC under light conditions (Imamura et al., 2003a). Furthermore, SigC contributes to global gene expression in the stadium of post-exponential growth, involving the central signal transducer PII (encoded by the nitrogen-related *glnB* gene) (Asayama et al., 2004).

Collectively, these observations and models raise the possibility that the cyanobacterial circadian timing process is not absolutely dependent on the *kai* gene expression cycle and that the "simple" cyanobacterial clock is more complex than was previously thought (Iwasaki and Kondo, 2004).

In addition to the standard group 1 factor (σ^{70}) that is responsible for transcription from a number of house-keeping promoters, all *Prochlorococcus* and *Synechococcus* strains sequenced so far harbour several more group 2 σ factors. In *Prochlorococcus* strains light/dark conditions induce cell synchrony (Jacquet et al., 2001). Distinct G1, S and G2 phases were found to characterise cell cycles of marine *Synechococcus* and *Prochlorococcus*. It was suggested that cell division in marine *Synechococcus* is controlled by circadian oscillators as it was described for *Synechococcus elongatus* PCC 7942 (Asato, 2003). For the sequenced marine strains, *Prochlorococcus* Med4, SS120, MIT 9313 and *Synechococcus* WH 8102, the existence of an internal autonomous circadian oscillator still awaits experimental testing.

1.7 The new era of small RNA molecules - tiny but mighty regulators

Small RNAs prove more influential than imagined: Short non-coding RNA molecules known as microRNAs and short interfering RNAs (siRNA) may control more than one third of our genes. This claim by Lewis et al. (2005) suggests that such RNA molecules could play a role in almost every process from cell birth to cell death.
However, small non-protein-coding RNAs (ncRNAs) with important regulatory roles are not confined to eukaryotes (Gottesman, 2005). Recent studies have shown that bacteria also possess a significant number of regulatory ncRNAs. These are a heterogeneous group of functional RNA molecules normally without a protein-coding function. They are frequently smaller than 200 nucleotides in size, and act to regulate mRNA translation/decay but can also bind to proteins and thereby modify protein function; for a recent review see Gottesman (2004).
It has been known for many years that non-coding RNAs control replication and maintenance of prokaryotic extrachromosomal elements (Wagner and Simons, 1994; Gerhart et al., 1998). A regulatory RNA, RNAIII, was found to play a critical role in *Staphylococcus aureus* virulence; reviewed by Johansson and Cossart (2003). More recently, up to

four different ncRNAs have been identified as crucial regulators in quorum sensing systems of *Vibrio* species, including the regulation of virulence in *V. cholerae* (Lenz et al., 2004). Moreover, ncRNAs are described as important factors in bacterial regulatory networks that respond to environmental changes (Sledjeski et al., 1996; Altuvia et al., 1997). Examples are OxyS: oxidative stress, DsrA: cold shock, IstR: SOS response, RhyB: iron stress and MicF: osmotic stress. All RNAs known so far are collected in the RFAM database including sequence and folding information as well as alignments with sequences of related species and software tools for the identification of new members (Griffiths-Jones et al., 2005).

The bacterial small RNAs include two major classes. The largest family acts by base-pairing with target mRNAs to modify mRNA translation or stability. For this mechanism, an RNA chaperone protein, Hfq, seems to be essential, which is similar structurally and functionally to eukaryotic Sm proteins: Hfq contains a conserved sequence motif, known as the Sm1 motif, forms a doughnut-shaped structure, and has RNA binding specificity very similar to the Sm proteins (Moller et al., 2002; Valentin-Hansen et al., 2004). It was demonstrated *in vitro* for OxyS (Zhang et al., 1998, 2002) and for another ncRNA, Spot 42 (Moller et al., 2002), that they act by pairing with complementary sequences in their mRNA targets (see also Fig. 1.9) and that Hfq is important for this pairing.

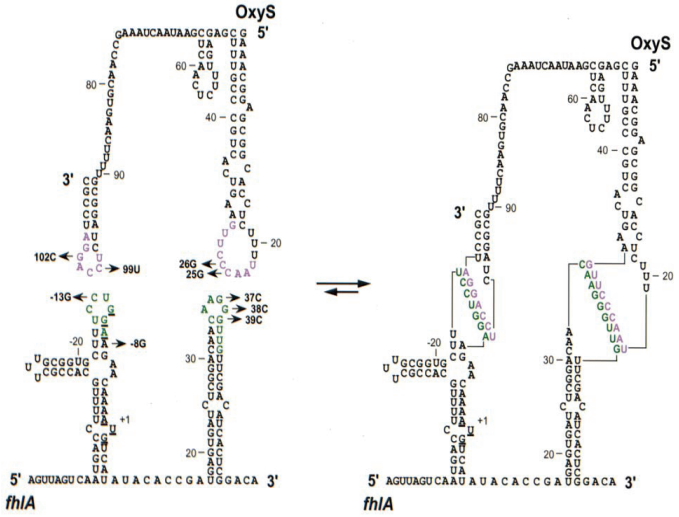


Figure 1.9: A model for *fhlA*-OxyS interaction, suggested by Argaman and Altuvia (2000). The Shine-Dalgarno and the initiation codon AUG are underlined. The numbering of *fhlA* starts at the initiation codon (AUG). The numbering of OxyS starts at the transcription start site. The arrows indicate point mutations introduced by Argaman and Altuvia (2000). The *fhlA* nucleotides of the Rbs (5' end stem-loop) and the Cds (3' end stem-loop) sites are in green and the complementary sites in OxyS are in purple forming a so called kissing complex by pairing with each other.

The second major category of regulatory RNAs are those that act by binding to a protein and modifying its activity. There are currently two known target proteins in *E. coli* regu-

lated directly by ncRNA binding: CsrA, the carbon storage regulatory protein, regulated by binding to CsrB and CsrC, two RNAs that inhibit CsrA (Weilbacher et al., 2003); and RNA polymerase, changed in promoter selectivity by binding to the 6S RNA (Wassarman and Storz, 2000).

Each ncRNA may regulate multiple target genes; for example OxyS acts as a global regulator on more than 40 genes in *E. coli*, including such important regulatory genes as *fhlA* (transcriptional regulator) and *rpoS* (alternative sigma factor) (Altuvia et al., 1998). The converse situation has also been found: Three different ncRNAs, OxyS, DsrA and RprA, converge on *rpoS* for regulation (Repoila et al., 2003).

1.7.1 How to find them?

Today, as a result of recent systematic searches, more than 70 ncRNAs are known in *E. coli*, most of which had been overlooked by traditional genome analysis. Thus, they represent about 2 % of the number of protein-coding genes, at the moment. Within a very short period, a variety of non-coding RNAs were found either by computational or experimental genome-wide screens. However, computational detection of ncRNAs is far more difficult than identification of protein-coding genes, because ncRNAs are typically short, have widely varying motifs and are often characterised more by their secondary structure than by their primary sequence (Huttenhofer et al., 2005). Some common properties are needed to serve as a guide for computational studies.
The first global searches in *E. coli* were based on the observation that known small RNAs mainly reside in intergenic regions and are often conserved in closely related species. Thus, searches were successful outside of protein-coding genes (Wassarman et al., 2001) combined with conservation of sequence characteristics expected for secondary structure (Rivas et al., 2001). The basic idea implemented by QRNA (Rivas and Eddy, 2001) and some other programs is that conserved structural RNA tends to show a pattern of compensatory mutations consistent with the base-paired secondary structure (stem-loop).
Additionally, intergenic regions were scanned for promoters and rho-independent terminators of transcription (a stem-loop followed by a row of Ts) with an orientation and spacing consistent with a ncRNA (Argaman et al., 2001; Chen et al., 2002). Several other computational studies exist, but not all have included experimental tests of their predictions (e.g. ddbRNA by di Bernardo et al. (2003)).
The identification of ncRNA genes by experimental methods can be achieved by generating specialised cDNA libraries of small transcripts (Vogel et al., 2003b; Kawano et al., 2005) or microarray techniques (Wassarman et al., 2001; Tjaden et al., 2002). Direct cloning of RNAs of a given size range may be useful for isolating ncRNAs expressed at a high level under a given growth condition or in the case that microarrays are not available (Gottesman, 2004). Small RNAs can also be identified by their association with a known RNA-binding protein like Hfq (Zhang et al., 2003).
Another promising approach are genome-wide searches for binding sites of known classes of regulatory proteins, which identified specifically regulated ncRNAs in *Pseudomonas aeroginosa* and *Vibrio cholerae* (Wilderman et al., 2004; Lenz et al., 2004; Gottesman, 2005).
All the successful approaches used over the years were reviewed recently by Vogel and

Sharma (2005),

- genome-wide searches based on the biocomputational prediction of ncRNA genes

- global detection of non-coding transcripts using microarrays

- shotgun cloning of small RNAs

- co-purification with RNA-binding proteins, such as Hfq or CsrA / RsmA

- classical cloning of abundant small RNAs after size fractionation in polyacrylamide gels

providing a nice overview about the state of the art of RNA finding.

1.7.2 Are the newly identified ncRNAs functional?

Both computational and expression-based methods are productive in finding small regulatory RNAs, but each has its specific limitations. Many ncRNAs have been already predicted and also verified, but for the majority of these RNAs little or nothing is known about function. Thus, further experimental and computational approaches are needed to determine whether these recently found transcription products are important parts of regulatory networks and biochemical pathways or simply 'junk' RNA transcribed as an unplanned by-product in the cell (Huttenhofer et al., 2005).

1.7.3 Riboswitches: mRNA-enslaved ncRNAs

New findings indicate that riboswitches are robust genetic elements that are involved in regulating fundamental metabolic processes in many organisms (for a review see Mandal and Breaker (2004)). Metabolite binding domains exist within several mRNAs that serve as precision sensors for their corresponding targets. An allosteric rearrangement of the mRNA structure (named riboswitch) is mediated by ligand binding, and results in the modulation of gene expression without the obligate involvement of a protein factor (Mandal et al., 2003); for an example see Figure 1.10. Numerous mRNAs in prokaryotes, including cyanobacteria, carry complex folded domains, which are known as riboswitches, within the non-coding portions of their polynucleotide chains.

TPP (THI element, Fig. 1.10) and cobalamin (B12-element (Griffiths-Jones et al., 2005)) are the two metabolite-binding riboswitches predicted for marine cyanobacteria, each with not more than one count per genome. One might expect to identify additional cases for genetic switches in these organisms by a more intensive, comparative search approach.

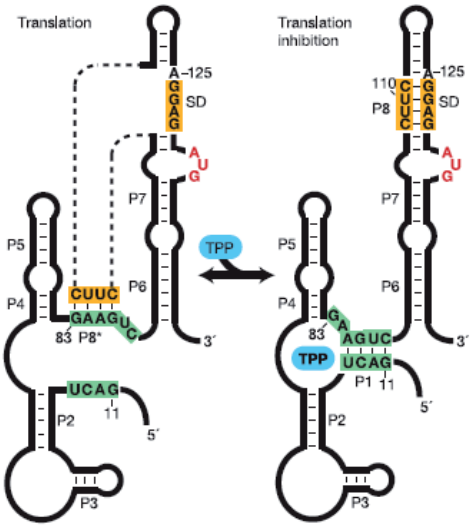


Figure 1.10: Schematic representation by Winkler et al. (2002) of the proposed mechanism for TPP-dependent deactivation of *thiM* translation. In the absence of TPP, the P8* pairing is formed between the anti-SD element and the anti-anti-SD element. This conformation permits the SD sequence to interact with the ribosome, and thus translation proceeds. In the presence of TPP (blue), the obligate formation of the P1 stem sequesters a portion of the anti-anti-SD element, and therefore the complete P8 stem also forms. This precludes ribosome access to the SD element, which inhibits translation. Complementary sequence elements that form P1 and P8 are depicted in green and orange, respectively (Winkler et al., 2002).

Two related and fascinating examples might be RNA thermosensors and T-box RNAs, gene-control elements that share some characteristics with metabolite-binding riboswitches (Mandal and Breaker, 2004). The 5' untranslated region (5' UTR) of the *pfrA* gene from *Listeria monocytogenes* carries a hairpin structure that occludes the ribosome-binding site. A shift in temperature from 30 to 37°C induces a fivefold increase in *pfrA* expression (Johansson et al., 2002), presumably by melting local base-pairs and permitting access by the ribosome. Thus, the RNA alone might serve as a molecular thermometer.
T-box RNAs are normally found in the 5' UTR of genes that encode aminoacyl-tRNA synthetases or related amino-acid-biosynthesis genes of gram-positive organisms. They fold into a structure that selectively recognises a specific tRNA (Grundy and Henkin, 2003), and activate gene expression in the presence of non-aminoacylated tRNAs, thereby boosting the expression of genes that are needed to maintain an adequate pool of charged tRNAs. A scenario that can be assumed for marine cyanobacteria as well.

1.7.4 Are there regulatory RNAs in cyanobacteria?

Many of these versatile bacterial riboregulators found in enterobacteria use base-pairing interactions to regulate the translation of target mRNAs. Because most of these antisense-acting ncRNAs have only incomplete target complementarity, duplex formation frequently depends on the activity of Hfq, an RNA chaperone. Only very recently, an *hfq* homologue

was predicted in cyanobacterial genomes, including two of the strains used in this study, *Synechococcus* WH 8102 and *Prochlorococcus* MIT 9313 (Valentin-Hansen et al., 2004). It supports the idea that riboregulatory processes similar to those of enterobacteria should exist in cyanobacteria as well.

At the begin of this work there was no information about the presence of regulatory RNAs and their genes in marine cyanobacteria. Apart from rRNA and tRNA genes, only three other well-characterised RNA genes were annotated by sequence similarity in each of the four marine genomes, *Prochlorococcus* Med4, SS120, MIT 9313 and *Synechococcus* WH 8102. These encode the RNA components of RNAse P (M1 RNA), the signal recognition particle (scRNA) and tmRNA (*rnpB, ffs* and *ssrA*, respectively). Although the *Prochlorococcus* tmRNA has not been analysed experimentally so far, it was subject to several *in silico* analyses, predicting it would consist of two separate molecules derived from a common precursor (Gaudin et al., 2002; Keiler et al., 2000). Such a permuted gene structure producing a two-piece mature tmRNA results in a dramatically reduced number of secondary structure elements. It remains unclear, if such a simplification in the structural elements of this RNA species would bring any selective advantage. However, the question raises, whether number and complexity of ncRNAs in these tiny, marine organisms is generally reduced as observed for tmRNA and regulatory proteins. And if that will be the case, what kind of ncRNAs might have escaped such an elimination and simplification process?

1.8 Aims of this work

The global importance of cyanobacteria belonging to the two genera *Prochlorococcus* and *Synechococcus* is indisputable these days. Environmental studies suggest that it is the specific niche adaptation into the complex and fluctuating oceanic environment, which has shaped the physiology and underlying genetic information of these organisms, driven their evolution and ecological success. Sequencing of multiple marine strains opened the great opportunity to study the molecular background of this exceptional adaptation process in a comparative genomics approach.

A successful comparative analysis, predicting conserved regulatory elements like DNA binding sites or secondary structures, strongly depends on a well-chosen set of sequences with reliable evolutionary distances. The marine genomes of *Prochlorococcus* Med4, SS120, MIT 9313 and *Synechococcus* WH 8102 might represent just such a perfect set.
Thus, the major aim of the present study was to analyse these marine genomes computationally, whereby the resulting predictions should be tested directly in the laboratory. Fortunately, all the four strains are growing in cell culture and for one marine strain, *Synechococcus* WH 8102, a genetic manipulation system was described previously, providing the minimum requirements for further biomolecular experiments.
In particular, experimental and computational promoter studies were aimed to get insights into the transcriptional regulation of cyanobacterial genes. For this purpose, phylogenetic footprinting algorithms has been successfully demonstrated to predict transcription factor binding sites within a given set of orthologous sequences. Additionally, the RACE

technique appeared very useful, which has been shown to determine sites of transcription initiation precisely.

Besides protein-coding genes, small non-coding RNAs possess a high regulatory potential, but systematic searches for ncRNAs are still lacking for most eubacterial phyla outside the enterobacteria. Thus, this work aimed at the question: Are small RNAs existing in marine cyanobacteria and if they are, might it be possible to derive hints on their function? Recently, an effective method to score multiple alignments in terms of secondary structure conservation was suggested (Washietl and Hofacker, 2004; Washietl et al., 2005). Using a comparative genomics approach based on the published genome sequences, one may predict candidates for ncRNAs in marine cyanobacteria. The expression of these candidate sequences needs to be tested under various growth and stress conditions that are encountered in the natural environment. The conservation of ncRNAs between different strains may provide information about an essential as well as a special ncRNA equipment for each strain with prospect to its niche adaption.

Another fascinating phenomenon proven for freshwater cyanobacteria is a unique circadian clock; otherwise found exclusively in eukaryotes but built of completely different components. The most studied clock model strain *Synechococcus elongatus* is the closest sequenced freshwater relative to the marine group - are they sharing a similar circadian pacemaker?

Chapter 2

Materials and Methods

2.1 Experimental Part

2.1.1 Cultures, cultivation, plasmids

Cultivation of marine cyanobacteria

Cultures of *Prochlorococcus* and *Synechococcus* were grown in a chemically-defined artificial sea water medium (ASW): *Prochlorococcus* Med4, NATL2A-MIT and *Synechococcus* WH 7803, RS9906, WH 8102. For cultures of *Prochlorococcus* SS120, MIT 9313, MIT 9312, MIT 9211, MIT 9215, which were not growing well on the first medium, PRO99 medium based on Atlantic seawater was used instead. The light conditions were 18 (Med4, 9312, 9215, WH 7803, RS9906 and WH 8102) or 10 (all other strains) μmol quanta m^{-2} s^{-1} white light at 23°C in a 12 h day - 12 h night cycle.
All cultures of marine cyanobacteria had originally been kindly provided by Dr. Frédéric Partensky (CNRS, Roscoff, France) or Dr. Penny Chisholm (MIT, Cambridge, U.S.) and existed at the begin of this work for at least two years in our laboratory.

Both media, ASW and PRO99, were prepared as described previously (Steglich, 2003; Chisholm, 2006). The seawater (artificial sea water for ASW: Milli-Q water with 0.48 M NaCl, 0.028 M $MgSO_4$, 0.027 M $MgCl_2$, 0.009 M KCl, 0.01 M $CaCl_2$; Atlantic seawater for PRO99) was sterile-filtered, autoclaved and allowed to cool overnight. Nutrient additions were given to the filtered and autoclaved seawater, resulting in the final concentrations: 0.05 mM phosphate buffer, pH 7.5; 0.4 mM $(NH_4)_2SO_4$; 1 mM HEPES, pH 7.5; 2 mM $NaHCO_3$.
Metal additions were again the same for both media, ASW and PRO99, and minor modifications resulted in the following final concentrations for trace metals in the medium: 1.2 mM EDTA; 1.2 mM $FeCl_3$; 8 nM $ZnCl_2$; 5 nM $CoCl_2$; 90 nM $MnCl_2$; 3 nM Na_2MoO_4; 9 nM Na_2SeO_3; 10 nM $NiCl_2$.

The use of 1 N HCl and Milli-Q washed and autoclaved glass-containers was essential for successful preparation of media. All nutrient and metal additions were prepared separately as primary stocks, sterilised using 0.2 μm VacuCap 90 filters (PALL, Dreieich, Germany) and stored at -20°C. The final media were stored at room temperature for up to one month. All marine cyanobacteria were cultured in 40, 160 or 400 ml polycarbonate

cell culture flasks (Nunc, Wiesbaden, Germany) as shown in Figure 2.1.

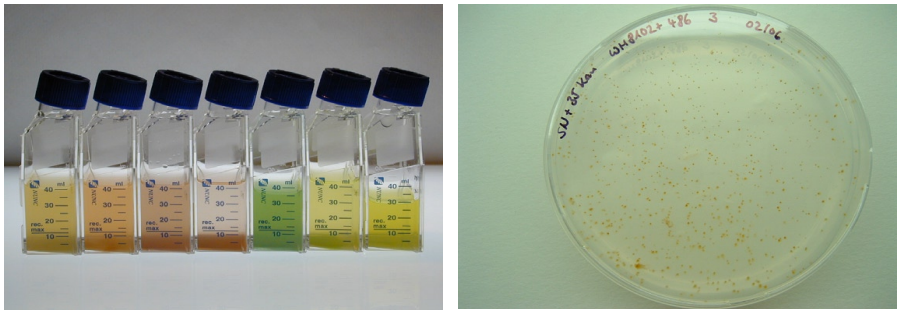


Figure 2.1: Different liquid cultures of *Prochlorococcus* or *Synechococcus* (left) and marine *Synechococcus* WH 8102 on plate (right).

Environmental perturbations for *Prochlorococcus* Med4

Prochlorococcus Med4 was subjected to various changes in growth conditions by individually depleting nitrate, phosphate or iron in artificial seawater; a shift from approximately 10 μmol quanta m^{-2} s^{-1} white light into darkness or into 10 μmol quanta m^{-2} s^{-1} blue light or into 50 μmol quanta m^{-2} s^{-1} daylight as high light condition, or the addition of DCMU to a final concentration of 2 μM for the inhibition of photosynthetic electron transport and the induction of severe oxidative stress; as well as temperature shifts to 15°C or 30°C. Med4 cultures were concentrated ten-fold by centrifugation for 10 min at 9,000 rpm at 15-20°C and cell pellets were washed once with the corresponding depleted media if necessary. The concentrated cultures were incubated for three hours at the respective condition.

SN medium for marine *Synechococcus*

For strains of marine *Synechococcus* optimum growth was observed during cultivation on liquid SN medium. This medium was described previously by Waterbury and Willey (1988); here Atlantic seawater was the basis.
Solid SN medium (Waterbury and Willey, 1988) was prepared by using ultrapure, low-melting point agarose (Invitrogen, Karlsruhe, Germany). For pour plating of single colonies, marine *Synechococcus* strain WH 8102 was serially diluted in liquid SN medium and 0.1 ml of the dilution was added to 35 ml of solid SN medium containing 0.6 % (wt/vol) low-melting agarose at 37°C and poured immediately into a plastic petri dish (Brahamsha, 1996). All plates were incubated at 23°C under 10 μmol quanta m^{-2} s^{-1} white light for 24 h and then moved to a higher intensity of 40 μmol quanta m^{-2} s^{-1} (Brahamsha, 1996). Colonies generally appeared from 10 to 14 days (Fig. 2.1), but it took up to six weeks before mutant colonies became visible. If incubated for more than two weeks, plates were wrapped in parafilm to avoid dry-out. Colonies were excised (at a

diameter of about 1 mm) and transferred to liquid SN medium.

Cryopreservation

For cryopreservation of marine cyanobacteria, 50 ml cultures were spun for 15 min at room temperature at 8,000 rpm. The pellet was resuspended in 1 ml original medium and transferred into a cryogenic vial (Nunc, Wiesbaden, Germany) where 7.5 % DMSO (75 μl) was added. The vial was inverted several times and placed immediately into liquid nitrogen and stored in a -80°C freezer.
Frozen cells were thawed in 37°C water bath until the half was liquid and transferred into sterile medium to revitalise the culture.

Other cyanobacteria and RNA samples

Cells of *Gloeobacter violaceus* PCC 7421 and *Synechococcus* PCC 7942 were ordered from Pasteur Culture Collection, Paris, France. *Synechocystis* PCC 6803 and *Anabaena* PCC 7120 were kindly provided by Dr. Annegret Wilde, Institute for Biology/Biochemistry, HU Berlin, Germany. These strains were grown in BG 11 medium (Rippka et al., 1979) under 30 to 40 μmol quanta m^{-2} s^{-1} white light at 23°C. For all other cultures, cell pellets or RNA samples were kindly provided by:

Dr. Jan Kern, Max-Volmer-Laboratory for Biophysical Chemistry, TU Berlin, Germany: *Thermosynechococcus elongatus* BP-1;

Holger Jenke-Kodama, Institute for Biology / Genetics, HU Berlin, Germany: *Nostoc punctiforme* ATCC 29133;

Dr. Elke Dittmann, Institute for Biology / Genetics, HU Berlin, Germany: *Microcystis* PCC 7806.

Dr. Debbie Lindell, MIT, Cambridge, U.S.: phage P-SSP7 RNA

Dr. Julia Holtzendorff, CNRS, Roscoff, France: RNA of synchronised *Prochlorococcus* Med4 culture

For *Synechocystis* PCC 6803 the accumulation of different transcripts from the *isiAB* region was studied under iron limitation and high light conditions. Therefore, liquid cultures of *Synechocystis* PCC 6803 were grown at 30°C in BG 11 medium (Rippka et al., 1979) under iron replete conditions and continuous illumination with white light of 50 μmol quanta m^{-2} s^{-1} to the logarithmic phase of growth. For high light conditions, cell cultures were exposed to white light of 170 μmol quanta m^{-2} s^{-1}. In parallel, logarithmic cell cultures were washed three times in iron-free media and grown further in iron deplete conditions

for several days.

Escherichia coli strains and plasmids

E. coli strain TOP 10 (Invitrogen, Karlsruhe, Germany) cells were used for preparation of eletrocompetent cells, transformation, blue-white colony screening and plasmid amplification according to standard protocols (Sambrook and Russell, 2001). Thereby, the pDrive cloning vector (QIAGEN, Hilden, Germany) was used for highly efficient cloning of PCR products.

The pUC19 vector (Fermentas, St. Leon-Rot, Germany) is a small, high-copy number *E. coli* plasmid which served here as a molecular marker after digestion with *Msp*I (*Hpa*II).

For genetic manipulation of marine cyanobacteria via conjugative biparental mating the plasmid pMUT100 (constructed by cloning the kanamycin-resistance cassette isolated from pUC4K into the *Pst* I site of pBR322 vector) as well as the helper system, *E. coli* MC1061 (pRL528, pRK24), were kindly provided by Dr. Martin Ostrowski, University of Warwick, GB. All these bacterial strains and plasmids needed for transformation were described previously by Brahamsha (1996).

2.1.2 DNA and RNA oligos

All DNA oligos were ordered at Metabion, Martinsried, Germany or Sigma-Aldrich, Steinheim, Germany. Their T_m and putative secondary structures were determined by the Oligonucleotide Properties Calculator (Cao and Kibbe, 2006). RNA oligos were delivered by Metabion, Martinsried, Germany. Each oligo is described below in the corresponding subsection.

2.1.3 Genetic manipulation of marine *Synechococcus*

Transconjugants of *Synechococcus* WH 8102 were obtained via conjugation, based on biparental matings as suggested by Brahamsha (1996). Here, the original protocol for conjugation was followed (Brahamsha, 1996), with the exception of the final plating step: After conjugation, the mating mixtures were cut out of the agarose with a sterile spatula and transferred into liquid SN medium without antibiotic for three days. After this additional incubation step, kanamycin was added (15 μg per ml first and up to 25 μg per ml after about one week of adaptation). At the earliest possible moment, mutant growth was observed on liquid medium with antibiotic, they were plated to obtain single colonies as described above.

2.1.4 Isolation of nucleic acids

Isolation of plasmid DNA from *Escherichia coli*

Plasmid DNA was isolated using the QIAprep Spin Miniprep Kit (tip 20 and tip 100; QIAGEN, Hilden, Germany) following the manufacturers instructions. The low-copy plasmid pMUT100 needed a special procedure described by QIAprep Midiprep for low-copy plasmids and cosmids in the manual.

Isolation of *Prochlorococcus* and *Synechococcus* DNA

The genomic DNA was purified as previously described (Franche and Damerval, 1988). A culture pellet from 400 ml original culture was resuspended in 2 ml ice-cold SET buffer. Addition of 1/4 vol 0.5 M EDTA and 20 % SDS (final concentration of 100 mM and 2 %, respectively) and approximately 0.5 mg proteinase K (Fermentas, St. Leon-Rot, Germany) resulted in cell lysis after incubation at 50°C for 2 to 16 h with gentle agitation. DNA was isolated by standard phenol-chloroform-extraction.

SET: 25 % sucrose, 50 mM Tris-HCl pH 7.5, 1 mM EDTA

Isolation of RNA

The cells were harvested by centrifugation for 15 min at 8,000 rpm and 4°C. Total RNA of marine cyanobacteria was isolated as previously described (Garcia-Fernandez et al., 1998). The pellet was resuspended in RNA resuspension buffer (0.5 ml for 400 ml culture), quick-frozen in liquid nitrogen and stored at -80°C until extraction. The frozen cell pellets were thawed under addition of 3 vol Z6 RNA extraction buffer (Logemann et al., 1987) for 30 to 60 min at room temperature and occasional shaking by hand. Acidic phenol (Roth, Karlsruhe, Germany) was added (0.5 vol) and the cells were incubated for another 60 min at room temperature, followed by phase extraction with 0.5 vol of chloroform:isoamylalcohol (24:1). After shaking, the phases were separated by centrifugation for 10 min at maximum speed. The aqueous phase was extracted then first with 1 vol phenol:chloroform:isoamylalcohol (25:24:1) and then with 1 vol chloroform:isoamylalcohol (24:1). Finally, the aqueous phase was transferred into a fresh tube, and the RNA was precipitated with 1 vol isopropanol for at -20°C overnight. After centrifugation at maximum speed for 30 min and 4°C the pellet was washed with 70 % ethanol, dried for about 10 min, resuspended in water and re-precipitated with 3 vol EtOH:NaOAc (30:1) overnight at -20°C. The RNA was pelleted again by 30 min of centrifugation, washed, dried and dissolved in water.
The lysis conditions for MIT 9313 and WH 8102 were modified as these strains gave poor RNA yields by the standard procedure. The resuspended cells from these strains were homogenised in Z6 buffer (Logemann et al., 1987) by several freeze-thaw cycles using liquid nitrogen over a time of 30 min, followed by the addition of 1 vol of acidic phenol and incubation at 60°C for another 30 min.

RNA resuspension buffer: 10 mM sodium acetate, 200 mM sucrose, 5 mM EDTA

Z6 buffer: 8 M guanidinium hydrochloride, 50 mM β-mercaptoethanol, 20 mM EDTA, 20 mM MES, pH 7 (NaOH) (Logemann et al., 1987)

NaOAc: 3 M sodium acetate, pH 4.8

To determine RNA stability in vivo, cells were treated with rifampicin, 200 μg/ml (SIGMA, Munich, Germany) and filtered rapidly (i.e. within 60 s) through Supor 0.45 μm membrane filters (PALL, Dreieich, Germany) at different time points after treatment. Filters were submersed in RNA resuspension buffer and quick-frozen in liquid nitrogen. RNA was isolated by dissolving the filter in acidic phenol at 60°C followed by the standard phenol-chloroform-extraction as described above.

An alternative protocol was used for the rapid isolation of many RNA samples in parallel. *Synechococcus* and *Synechocystis* cell pellets were resuspended in Trizol (Invitrogen, Karlsruhe, Germany) and frozen at -20°C. The tubes were transferred directly from the freezer to a 65°C waterbath for 10 to 15 min, vortexing from time to time for lysis. The phase extraction was done in phase lock gel tubes (Eppendorf, Wesseling-Berzdorf, Germany) by adding 0.2 ml chloroform per ml of Trizol and by shaking for 15 s and resting at room temperature for 3 min. The following centrifugation step was lasting for 12 min at maximum speed which separated the phases. The upper aqueous phase which contains the RNA was transferred to a new tube and precipitated by adding 0.5 ml of isopropanol per ml Trizol used in the initial homogenisation.

Cyanobacterial cells such as *Nostoc punctiforme*, harbouring a very robust cell wall, were disrupted by adding an equal volume of glass beads, 33.3 μl 20 % SDS solution and 583 μl acidic phenol to 500 μl concentrated cell solution, following several cycles of vigorous agitation, freezing in liquid nitrogen and thawing in a waterbath. The centrifugation of the mixture for 15 min at maximal speed at 4°C yielded an upper aqueous phase which could be cleaned up by standard phenol-chloroform-extraction as described above.

Determination of concentration of nucleic acids

The concentration of nucleic acids was determined spectrophotometrically by measuring the extinction (E) at the wavelength of 260 nm. An optical density (OD_{260nm}) of 1 corresponds to 50 μg/ml dsDNA or 40 μg/ml ssRNA and 30 μg/ml of oligonucleotides. Proteins absorb light of a wavelength of 280 nm, therefore the ratio of E260:E280 can be used as an indication of the purity of the DNA. The ratios should be between 1.8 and 2. The quality and quantity of nucleic acids was further examined optically in an ethidium bromide stained agarose gel.

2.1.5 DNA gel electrophoresis, Restriction, Ligation

Double stranded DNA molecules of 80 bp to 10 kb were separated on 1.8 to 1.0 % agarose gels containing ethidium bromide (0.2 μg/ml gel) in 1 x TAE electrophoresis buffer after addition of 1/6 volume of 6 x DNA loading buffer under a field strength of 5 to 10 volt/cm. For DNA fragments smaller than 80 bp, 3 to 4 % NuSieve GTG agarose (Cambrex Bio Science, Verviers, Belgium) could resolve fragments differing by as little as 1 bp. After migration, the DNA was visualised on an UV transilluminator, gel imager (BIO-RAD, Munich, Germany). The standard markers pUC19/*Msp*I (*Hpa*II) and λ/*Pst* I were used to compare the sizes of the loaded DNA molecules.

6 x DNA loading buffer: 50 % (w/v) glycerol, 1 mM EDTA (pH 8.0), 0.005 % bromphenol blue and 0.005 % xylene cyanol

50 x TAE buffer: 2 M Tris, 2 M boric acid, 100 mM EDTA, pH 8.3 (Sambrook and Russell, 2001)

0.1 to 10 μg DNA were incubated with 5 to 10 U restriction enzyme (Fermentas, St. Leon-Rot, Germany) in a final volume of 10 to 40 μl using the provided buffers at the optimal temperature for the enzyme.

100 ng linearised vector was incubated with a 3 fold molar excess of insert, 1 x T4 ligase buffer and 1 U T4 DNA ligase (Fermentas, St. Leon-Rot, Germany) in a final volume of 10 μl at 4°C overnight, or 3 hours at room temperature, or for 1 hour at 37°C.

2.1.6 RNA gel electrophoresis, Northern blotting and Hybridisation

Agarose gel electrophoresis and capillary blotting

Total RNA was separated on denaturating 1.5 % agarose gels. The agarose was first dissolved in boiling water and cooled down to 60°C until addition of 0.1 vol 10 x MEN-Puffer and 16 % (vol/vol) formaldehyde. Prior to starting the gel, samples were mixed with 1 vol RNA loading buffer and denatured for 10 min at 65°C and kept on ice until loading. Migration occurred under a field strength of 10 volt/cm in 1 x MEN as electrophoresis buffer. For size-comparison, the standard RNA ladders High Range and Low Range (Fermentas, St. Leon-Rot, Germany) were included on the gel.

2 x RNA loading buffer: 95 % deionised formamide, 0.025 % (vol/vol) SDS, 0.5 mM EDTA, 0.4 % ethidium bromide

10 x MEN: 200 mM MOPS, 50 mM NaOAc, 10 mM EDTA, pH 7 (Sambrook and Russell, 2001)

The gel-separated RNA samples were blotted with 10 x SSC buffer to a Hybond-N+ membrane (Amersham, Freiburg, Germany) by capillary blotting; UV-crosslinking and

stripping (hot SDS procedure for re-use) as described in the manual of the membrane.

20 x SSC buffer: 3.0 M NaCl, 0.3 M sodium acetate, pH 7 (Sambrook and Russell, 2001)

PAA gel electrophoresis and wet blotting

Total RNA was separated in 10 % polyacrylamide-urea gels, 15 x 20 cm, in a vertical gel electrophoresis apparatus, Protean II (BIO-RAD, Munich, Germany). Electrophoresis conditions were about 5 volt/cm field strength for about 20 hours in 1 x TBE. Polyacrylamide gels were stained with ethidium bromide (0.3 μg/ml) in 1 x TBE buffer, rinsed with water and analysed with a Lumi-Imager F1 system (Roche, Mannheim, Germany). Transcript sizes were determined by correlation to *Msp*I-digested DNA of plasmid pUC19 (Fermentas, St. Leon-Rot, Germany), which was radiolabelled like oligonucleotide probes, see below. Gel-separated RNAs were transferred to Hybond-N+ membranes (Amersham, Freiburg, Germany) by electroblotting in a wet blot tank (BIO-RAD, Munich, Germany) overnight, following the manufacturers instructions; UV-crosslinking and stripping (hot SDS procedure for re-use) as described in the manual of the membrane.
Three different ways were used to verify that the same amounts of RNA samples were loaded in Northern blots: first, by measurement of RNA concentrations; second, by direct comparison of rRNA band intensities after staining by ethidium bromide; and third, by control hybridisations using the 5S rRNA as an internal standard.

loading buffer 2 x RPA: 98 % deionised formamide, 2 mM EDTA, 0.1 % xylene cyanole, 0.1 % brom phenol blue

10 x TBE buffer: 8.9 M Tris, 8.9 M boric acid, 200 mM EDTA (Sambrook and Russell, 2001)

10 % PAA urea gel: 10 % Rotiphorese acrylamide/bisacrylamide stock solution 40 % (Roth, Karlsruhe, Germany), 8.3 M urea, 1 x TBE, APS, TEMED (Sambrook and Russell, 2001)

Hybridisation

Following prehybridisation for at least 10 min in prehybridisation buffer at 45°C, oligonucleotide probes, which were 5' labelled by polynucleotide kinase (Fermentas, St. Leon-Rot, Germany) with 30 μCi γ^{32}P-ATP (Amersham, Freiburg, Germany), were added and usually hybridised at 52°C (depending on T_m of the oligo) for at least four hours. All DNA oligonucleotides are listed in Tables 2.1 and 2.2. The membranes were washed in washing solution I at 45°C for 10 min; washing solution II at 45°C for 5 min; and shortly in washing solution III at ambient temperature. Signals were detected and analysed on a Personal Molecular Imager FX system with Quantity One software (BIO-RAD, Munich, Germany).

name	sequence (5' to 3')	hybridisation to
ScoccusREV	GCGACGCCGTTTTACCT	6Sa RNA in *Thermosyn., Nostoc punctif.* and PCC 7942
PCC6803REV	CACCACGCCGTTTTACCT	6Sa RNA in PCC 6803
NosGloeoREV	CGCAACGCCGTTTTACCT	6Sa RNA in PCC 7120 and *Gloeobacter*
isrRrev	GGCAGGCAAAGCCATGTATTGGGGG	IsrR in PCC 6803
PMM3822n	TCCCTGGTATGACTCGAACTGC	mRNA of PMM3822n in Med4
5S	GGCATTGAGCTATTTTCTCAGGGGGCT	5S rRNA
t_ser	CGGAGAGGGAGGGATTCGAAC	tRNA-Ser
tm5	CGGAATCGAACCGCTGTCCGA	tmRNA (5'end)
tm3MITWH	TGACGGGAGAAACGAACGATGTTGT	tmRNA (3'end) in MIT9313 and WH8102
tm3MS	GCTGTTTGACGGGAGAAACTAACGAT	tmRNA (3'end) in Med4 and SS120
sc	CCTTCCGRTCCTGACCAGRTTT	scRNA
y1M	CGATTTGGTGTGTGAGGAGTATGG	Yfr1 in Med4
y1MIT	CGGTGTGTGTGAGGAGTGTGT	Yfr1 in MIT9313
y1WH	GGGGTGTGAGGAGTGTGTGG	Yfr1 in WH8102
y_genM	CCACTGTTTCAGTAAAACCTCTCCTAC	Yfr2-Yfr5 in Med4
y2aM	ATAGAGGTCTTTCCTTGGTTTAA	Yfr2 Med4
y3aM	ATTTCTAGAGATCTTTATATAGTTTTACT	Yfr3 Med4
y4aM	TCTAGAGGTCTTTTTCTGGTTTT	Yfr4 Med4
y5aM	AAGGGGCCTAAATTGGTTCTA	Yfr5 Med4
y2to5M	GTGTTTCCTTGTTTCCACTG	Yfr2-Yfr5 in Med4
y2S	GTGTTTCCTAGTTTCCACTTTT	Yfr2 in SS120
y2MIT	GTGTTTCCTTGTTTCCACTC	Yfr2 in MIT9313
y6MS	GTTGAACGTTTTRGTAGCKGTTGCTAC	Yfr6 in Med4 and SS120
y7MS	CCTCAAGTCGAAAAAGAGTCAGATCAGA	Yfr7 Med4 and SS120
y7MS_2	GTCGAAAAAGAGTCAGATCAGAGCAC	Yfr7 Med4 and SS120
y7MIT9312	GTCGAAAGAAAGTCAGATCAGAGCAC	Yfr7 MIT9312
y7MITWH	TCGAAAGAGAGTCAGATCAGAGCAC	Yfr7 MIT9313, WH8102 and WH7803

Table 2.1: DNA oligos used for PNK labelling and hybridisation to Northern blots.

For hybridisation of mRNAs, the HexaLabel DNA labelling kit was used (Fermentas, St. Leon-Rot, Germany) by following the manufacturers instructions. Therefore, the labelled triphosphate α^{32}P-dCTP (Amersham, Freiburg, Germany) was chosen. The DNA template (about 350 bp in length) for the radioactive DNA labelling was obtained by standard PCR, described below. The steps and solutions of the hybridisation procedure were the same as for oligonucleotide probes, described above. Only hybridisation and washing temperature were slightly increased to about 55°C and 50°C, respectively.

prehybridisation buffer: 50 % deionised formamide, 7 % SDS, 250 mM NaCl, 120 mM Na-PO_4, pH 7.2

washing solution I: 2 x SSC, 1 % SDS

washing solution II: 1 x SSC, 0.5 % SDS

washing solution III: 0.1 x SSC, 0.1 % SDS

name	sequence (5' to 3')	no signal to
ScoccusFW	AGGTAAAACGGCGTCGC	6Sa RNA in *Thermosyn., Nostoc punctif.* and PCC7942
PCC6803FW	AGGTAAAACGGCGTGGTG	6Sa RNA in PCC6803
NosGloeoFW	AGGTAAAACGGCGTTGCG	6Sa RNA in PCC7120 and *Gloeobacter*
isrRfw	GGATCTAATGTTGGCTGACTGACTAG GCG	putative complement IsrR-sequence in PCC6803
y1S	TTCTGCAGCAATCTAAATTTTTAAGA GAAGAAAAATAA	*guaB-trxA* intergenic region for the presence of Yfr1 in SS120
ID273MSt44rev	CCCGAAACAGTCAGATGTG	putative t44 RNA in Med4 and SS120
ID273WHt44rev	CCCGAAACGGGCAGG	putative t44RNA in WH8102
ID273MITt44rev	CCCGAAACGGACAGGT	putative t44RNA in MIT9313
ID217Mrev	CAGGGGTTGATGGGTC	cand. RNA CLID217 in Med4
ID217Mfw	GACCCATCAACCCCTG	cand. RNA CLID217 in Med4
ID102Mfw	CCACAGAAACTACTACTACTATATAA ATTCATTATTAG	cand. RNA CLID102 in Med4
ID102Mrev	CTAATAATGAATTTATATAGTAGTAG TAGTTTCTGTGG	cand. RNA CLID102 in Med4
ID185Mfw	AGATCGTATTCATCCATGGTTAAGTA AAGTTAAA	cand. RNA CLID185 in Med4
ID185Mrev	TTTAACTTTACTTAACCATGGATGAA TACGATCT	cand. RNA CLID185 in Med4
ID66Mfw	TATGAGTGAATATGATAATGAATATC AATAGCAATATCAA	cand. RNA CLID66 in Med4
ID66Mrev	TTGATATTGCTATTGATATTCATTAT CATATTCACTCATA	cand. RNA CLID66 in Med4
ID194Mfw	GTAGGAGAGGTTTTACTGAAACAGTGG	cand. RNA CLID194 in Med4
ID257fw	CCCCATACTCCTCACACAC	cand. RNA CLID257 in Med4
ID53Sfw	GTAGCAACAGCTACTAAAACGTTCAAC	cand. RNA CLID53 in SS120
ID53Mfw	GCAACCGCTACCAAAACGT	cand. RNA CLID53 in Med4
ID51Sfw	GTCATGCAACTGGAAATGGATCATTAC	cand. RNA CLID51 in SS120
ID51Mfw	GTAATTGGAAACGAAAGCCCA	cand. RNA CLID51 in Med4

Table 2.2: DNA oligos used for PNK labelling and hybridisation to Northern blots, which did not yield signals.

2.1.7 Polymerase chain reaction (PCR)

DNA fragments were amplified by Taq DNA Polymerase (QIAGEN, Hilden, Germany) in a standard PCR. A typical PCR cycle consisted of an initial denaturation at 93°C for 2-3 min followed by 34 three-step-cycles of denaturation at 93°C for 10 s, annealing at about 60°C (depending on the primer set) for 30 s and elongation at 72°C for about 20 s (depending on the length of the amplified fragment). The final extension at 72°C for 5 min terminated the PCR. A list of all used PCR primers is given in Table 2.3.

The primer set, pDriveFW and pDriveREV, anneals on vector pDrive and was utilised for amplification of the ligated insert, which was mainly used for screening of *E. coli* colonies ('colony' PCR).

For PCR systems that did not work well under standard conditions, a special protocol was adapted, described in the manual, including Q-solution (QIAGEN, Hilden, Germany) and an elongation at 72°C for 1 min per 1 kb amplified fragment. The Q-solution changes

DNA oligo name	sequence (5' to 3')	organism/plasmid
pDriveFW	ACGACGTTGTAAAACGACGG	pDrive
pDriveREV	TTCACACAGGAAACAGCTATGAC	
WHYfr1suiSTARTfw	GGTATGGCGACACCAAGCCCTGTC	*Synechococcus* WH8102
WHYfr1suiENDrev	TGGGCATTGCCCGCACACTCTTCACA	
WHYfr7suiSTARTfw	AATCCCCACTTTGTTTCCCCTTCTGGTTCTTC	
WHYfr7suiENDrev	TGGAACAGAGTCAGCTTCAAAAGGTCGTCG	
WHYfr7suiIN197fw	GCCCGATGACGCTCAGGTTGCTTG	
WHYfr7suiIN177rev	CAAGCAACCTGAGCGTCATCGGGC	
Yfr7colonyPCRfw	ATCCTGAACTGCTGTGTTGGTGACGGC	
Yfr7colonyPCRrev	TAAGAGGTCGCAGCCAGACGCTTCAAC	
Kan782FW	CGCCTGAGCGAGACGAAATACGCG	pMUT100, pUC4K
Kan1541REV	ACCTTCTTCACGAGGCAGACCTCAGC	
isiAsondeFW	CTGGGGCTTTGTTTCATACC	*Synechocystis* PCC6803
isiAsondeREV	TAAAATTCTGGCGGATAGGC	
isi-5utr-fw	TTGGGCGATCGCCAAAAATC	(Vinnemeier et al., 1998)
isi-5utr-rev	CTCTGTCCACCCGAGACCTA	(Vinnemeier et al., 1998)

Table 2.3: PCR primer list.

the melting behaviour of DNA and can improve the PCR result.

2.1.8 Sequencing of DNA

The 'colony PCR' fragments were purified on QIAquick spin columns (QIAGEN, Hilden, Germany) and prepared for sequencing by adding water and 10 pMol primer pDriveREV. Sequencing was performed by Dr. Martin Meixner (DLMBC, Berlin, Germany).

2.1.9 Rapid amplification of cDNA ends (RACE)

For RACE experiments, the RNAs were digested with RNAse-free DNAse I (Ambion, Austin, Texas USA). 30 μg RNA were incubated for 15 min with 6 units (3 μl) DNase I, 1 x DNase buffer and 40 units (1 μl) RNase inhibitor (Fermentas, St. Leon-Rot, Germany) at 37°C in a final volume of 40 μl. The DNase I was inactivated by incubation for 3 min at room temperature with 4 μl of the provided glassbeads-based inactivation reagent. After centrifugation, the RNA remained in the aqueous phase and was transferred to a fresh tube.

Mapping of RNA ends was performed by rapid amplification of cDNA ends as described previously (Bensing et al., 1996; Argaman et al., 2001) to determine experimentally the sites of transcription initiation and termination in cyanobacteria.

5' RACE
Primary transcripts in bacteria carry a 5' triphosphate, which can be cleaved specifically by tobacco acid pyrophosphatase, TAP (Epicentre, Madison, Wisconsin USA). The resulting 5' monophosphate was subsequently ligated, using T4 RNA ligase (Epicentre, Madison,

gene	RT primer (5' to 3')	PCR primer (5' to 3')
isrR6803	GGGCCGGAGCTCTAC	GGCAGGCAAAGCCATGTATTGGGGG
		CCCATTTTGCCAGCATCAGTAGCCTAGAAG
ssaA6803	CACCACGCCGTTTTACCT	GGTGTAACCAGGGGGGAGTCATCAATTC
		CTGCCATTATAACCCCATCCTTCGGCG
ssaA7120	CGCAACGCCGTTTTACCT	CCAGGTGGGTACCGATTTTCCTCGTTTAAAG
		CTCACTCATTATAGCTTTCAAAAACAGGTTCTACC
yfr1Med	ACAGACTTAAAAAAGCCCGATAA	GCGATTTGGTGTGTGAGGAGTATGGGG
yfr2Med	ATAGAGGTCTTTCCTTGGTTTAA	CTAAATCAAGTGTTTCCTTGTTTCCACTGTTTCAG
yfr3Med	ATTTCTAGAGATCTTTATATAGTT TTACT	AACCAAGTGTTTCCTTGTTTCCACTGTTTCAGTAAAA
yfr4Med	TCTAGAGGTCTTTTTCTGGTTTT	CCGAATCAAGTGTTTCCTTGTTTCCACTGTTTC
yfr5Med	AAGGGGCCTAAATTGGTTCTA	CGAGTGTTTCCTTGTTTCCACTGATTTAATAAAACC
yfr6Med	CTGTATGCTGTTTTTTTGCAGC	GTTTTGGACGTATTTTTGTCCTCTGTAGATAAGTGTCATC
yfr6SS120	CATGCTTCTTTTCAACTACTGC	GTTGAGTGTACTTTTGACCTCTGTAGGTTAGTGAC
yfr7Med	AAGCCAGAAGAGGAGTCAA	GCTTTCGTTTCCAATTACAGGATTGGTTTACGTCTACAT
		CACATAAAGAAGAACTTGACCAAAGATCCAGTATGTAAATA
yfr7SS120	AAGCCAGAAGAGGAGTCAA	TGATCCATTTCCAGTTGCATGACTGGTTTACTTCAC
		CTTAAAAGGTCATCACCAACAAAACAGTTCTCTTGCATA

Table 2.4: DNA oligos used for 5' RACE experiments analysing ncRNAs of *Synechocystis* PCC 6803, *Anabaena* PCC 7120, *Prochlorococcus* Med4 and SS120.

Wisconsin USA), to the 3' hydroxyl group of a RNA-oligonucleotide (5' adaptor: GAU AUG CGC GAA UUC CUG UAG AAC GAA CAC UAG AAG AAA), which was followed by reverse transcription with a gene specific oligo (placed within the first 200 nt downstream a start codon) and PCR amplification with a 5' adaptor (ATA TGC GCG AAT TCC TGT AGA ACG AAC ACT AG) and a nested gene specific primer (listed in Tab. 2.4, 2.6 and 2.5).

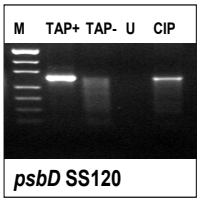


Figure 2.2: PCR result of 5' RACE for *psbD* in *Prochlorococcus* SS120. A single prominent band was obtained with TAP-treated RNA (TAP+). The presence of a weaker band of the same size in the control reaction (TAP-) indicates that a minor proportion of the primary transcripts were present in 5' monophosphate form in the original transcript population. Additional controls are RNA without any treatment (U) and 5' dephosphorylated RNA using calf intestine phosphatase (CIP). The standard marker pUC19/*Msp*I (M) was used to compare the sizes of the loaded DNA molecules.

TAP-treatment is expected to yield a specific or at least strongly enhanced signal (lane TAP+ in Fig. 2.2) for primary transcripts in the amplification step, as compared to untreated RNA samples (lane TAP- in Fig. 2.2). Because of this selectivity for newly initiated transcripts among the pool of 5' ends, and the small amounts of RNA material required, this technique was considered more suitable for this approach than traditional primer extension analysis.

Altogether, three controls were included in all 5' RACE experiments: RNA mock-treated

gene	RT primer (5' to 3')	PCR primer (5' to 3')
P001	TCTCCATAATTGACCCGCTT	CTTAATAAAGTCAGTCAGTCCATCCCAGTCAGT
P003	TCTCCCCCTTACCCATAG	CCTTCTTACCGAATTTTCCTATGGTGTACTTAG
P004	TTGTGCTTGGGTTTCTCTCA	GGAATCGTTTACAGGGAATAACTCAGGCTCG
P013	TCTTTACGTGGGGAGAATAG	GGTTATCTTTGCAGTTATAGCTCCTTGCGATTCTG
P029	GCCTTGTCAGCATTACCC	GGAAGGAACTTGTCATACGACCTGTGTAG
PMM1500	GATGAGATTTGACACCCTC	CCATCCATATCCTTGCTCTGGGCCG
PMM1501	GGACCTAGATCTGATACATG	CTATAAAGGCAGCATCAATACCTGGTAGGAC
PMM0684	TTGTAAATAAGAGATGGGTATAAAC	CTGCCTTGCGGGTTAAGTATCCAGCC
PMM0819	CATTTATCCCAGGCTTTTTCC	TGACTTGATGACTTGATTGCTGAGGGCTC

Table 2.5: DNA oligos used for 5' RACE experiments for cyanophage P-SSP7 infecting *Prochlorococcus* Med4.

in TAP buffer but omitting TAP (TAP-), RNA without any treatment (U) as well as 5' dephosphorylated RNA (CIP) using calf intestine phosphatase (Fermentas, St. Leon-Rot, Germany), see Figure 2.2.

3' RACE
3' hydroxyl ends of the RNA samples, either 5' blocked by 5' adaptor ligation or by dephosphorylation with calf intestine phosphatase (Fermentas, St. Leon-Rot, Germany), were ligated to the 5' monophosphate end of another RNA-oligonucleotide (3' adaptor: pAAG AUG AAU GCA ACA CUU CUG UAC GAC UAG AGC A ddC). Reverse transcription of the ligated RNA molecules was initiated with a 3' adaptor-specific oligo (GTG CTC TAG TCG TAC AGA A), followed by PCR amplification, each with a gene specific (for *isrR* of PCC 6803: GGA TCT AAT GTT GGC TGA CTG ACT AGG CG) and the 3' adaptor primer (GTG CTC TAG TCG TAC AGA AGT GTT GCA TTC ATC). Using the products from this PCR as a template, a second, nested PCR step was required to obtain distinct PCR products. In this nested PCR, for each gene a nested gene specific primer (for *isrR* of PCC 6803: CCC CCA ATA CAT GGC TTT GCC TGC C) and the 3' adaptor primer was employed.

Upon cloning and sequencing of the amplification product, the first nucleotide downstream of the 5' adaptor RNA or the last nucleotide upstream the 3' adaptor was assigned the transcription initiation or termination site, respectively.

2.1.10 Quantitative RT-PCR

DNase conditions
For quantitative RT-PCR analysis, the RNAs were digested with RNAse-free DNAse I (Ambion, Austin, Texas USA). 3 μg RNA were incubated for 20 min with 6 units (3 μl) DNase I, 1 x DNase buffer and 40 units (1 μl) RNase inhibitor (Fermentas, St. Leon-Rot, Germany) at 37°C in a final volume of 30 μl. The DNase I was inactivated by incubation for 3 min at room temperature with 3 μl of the provided glassbeads-based inactivation reagent. After centrifugation, the RNA remained in the aqueous phase and was transferred to a fresh tube.

Reverse transcription

gene	RT primer (5' to 3')	PCR primer (5' to 3')
trxAMed	ACTGCTCCAACAACAGTATC	CTGATTCCATATTGACTTGCAACATTTGGATTCTCATC
trxASS120	ATCATTAATGTAGGGATGCTTC	GCCATATTGACTTGCAACATTGGGATTCTCATC
		ATGGCAAGAATGATTTTTTTCGCCAAAATAAATAAGAGCTG
trxAMIT	TGGCTTTGGGCACTGCA	CAGGGTGGGGATACTGCGAATACCG
		GGTCGTTGATTTTGCGGAAATTTAGCAATCGTTTG
trxAWH8	CGTTGGGATGCTGCGG	CCGTACTGACTGGCGACATTAGGGTTTTC
pcbASS120	AGTAGCTAAGTGAGGAAGAG	GTACTGAAGGGTCAAAACGAGCTAATTCAAAAAGTGTGAA
		GGGAAAGCGCATAAACACTAACGAAAAGAGCGA
psbASS120	TGGGTAGAAGTGAAGACCTAT	GCACCTGAAATGATGTTGTTCCCATAAAGGAAAGATC
		CCGCCAAGAGGAGTTTAGATGAAAGTTTGCAA
psbDSS120	GACTGTGCCCCATAGC	GGGTACTAACAGCGGCAGTAAGGAAATTACATC
		GGATACGCCTGAGACTTTAACAATTCAAGAACCAAG
petHSS120	TGTTTACAGGAACATTGGGGT	ATTGCCTTTGGCGAGGAAGAAGGCTTTTCAG
		CCAATATCTCAACTTTATTATTCGACATAGCCTTTTTCAAAG
kaiBSS120	GGGGTTGCAAGTATCTTATCT	GCTAGCTGTGGGTTTTTAAGCACATCAATTACTTTCAAAG
		TCTAAATCTTGACGGATTTAACCTCTAAGAGTGAGGATG
ftsZSS120	CTAGGCCTCGAGTTAATGT	GAACTCTATTCTCAGCCGAAGATTGCAACAAAGC
		GGTGCATTATGACTGGCTTTGGAGATCTGTTG
ntcASS120	GTCGAACAGCTCCTCTAC	GCAGGGTCCCCGGGGAAAAAGATC
		CCAATTCAAAGGACGATAAGAAGTGAAGCTAAAACC
cpeBSS120	CTTCTAGTGGGATAGCAATG	CTCACAAACCATTCCTGATATAGCATCTGCAGC
lexAMed	TTCAATTAAGCCACCAGCTG	CCCATAATTGGGACTCCCTCAAAGATCTCATC
umuDMed	CTGGTTTAGCAGTTAGACTC	CTAAGAAAAAAAGTACTAAAAGGATTAGATATTAAATGCT
		CGTTC
PMM1427	AGCTGCTTCTGGAGCAAATA	CGACTATCTTTCCTCCAGTAAGTACAGCCC
glnAMed	GAATTCTGGCTCTGGTCC	AGGGTTCTCCACTTCTTGGCTCTTGAATAG
urtAMed	TGTCCATCCACCAAATACAAC	CTCAGCAAATGTTGGCCAGTCGGAAGC
psbDMIT	GCTATGACCCATCGCATC	CTTACAGCGGCGCTAAGGAAATTGCAGC
psbDWH8	CTGTGACCCATCGCATC	AGACAGCAGCGGTGAGGAAGTTGCAAC
psbDWH7	CTGTGACCCATCGCATC	GGTGGACACAGCAGCGGTGAGG
atpBWH7	CGAGAACGTTGAAGATTCG	GTTGCTTCACCCACGGGAACGCTG
rpl21WH8	ATGGGCCATCACCTTCAG	CTTGGAATCCTTCACCATCAGAACGTTTTCG
rpl21MIT	GTGGTCCATCACCTTCAG	CTTGCCGTCTTTGATCAGCAGTACCTTCTC
		CCGCGTTTAGGCGCCTTTAGTCAAGATTG
rpl21SS120	ATGGATGTAACCATTACTCTTGT	CTCTTGCCTGTGCCCATTCTTTGTGCG
		CAAGATAACATCCTCACTCTTAGAGGTTAAATCCG

Table 2.6: DNA oligos used for 5' RACE experiments analysing mRNAs of *Synechococcus* WH8102 (WH8), WH7803 (WH7), *Prochlorococcus* Med4 (Med), SS120 (SS120) and MIT9313 (MIT).

The First-strand cDNA Synthesis System for Quantitative RT-PCR (Marligen Biosciences, Ijamsville, MD U.S.) was used to transcribe about 300 ng DNA-free RNA into cDNA, initiated with oligo $(dT)_{20}$ and random primers, following the manufacturers instructions.

Primer design

Prochlorococcus Med4 gene-specific RT-PCR primers for the gene *rnpB* as well as for *isiB, glnA, pstS* and *hli8* were designed, based on the genome sequence using PrimerExpress software (PE Applied Biosystems, Weiterstadt, Germany). The following parameters were required: length of the PCR product: 80 to 120 bp, length of the primers: 20 to 25 nt, melting temperature of the primers: 58 to 62°C and GC content: app. 40 %. A list of the designed primer sets is given in Table 2.7.

Quantitative PCR

For PCR, 1 μl of the reverse transcription reaction product was used in a 25 μl reaction with 1.8 pmol of each primer, 1 x SYBR Green buffer, 0.2 mM each dNTP, 3 mM $MgCl_2$, 0.625 U Ampli Taq Gold Polymerase (PE Applied Biosystems, Weiterstadt, Germany). PCR was performed in triplicates in a GeneAmp 7500 real time thermocycler (PE Applied Biosystems, Weiterstadt, Germany) with one initial denaturation step at 95°C for 10 min

gene	FW primer (5' to 3')	REV primer (5' to 3')
rnpB	TTGAGGAAAGTCCGGGCTC	GCGGTATGTTTCTGTGGCACT
isiB	GATGAGGAAAGGTCTGGAACTG	TTCTGTATAGGTAGATGAATCCCC
glnA	GGAATACTGGTCGAACAGAAGAA	CCTTATGTCTTGTGCGGTGTC
hli8	CTCTAACTGCTTTACTCGTTGC	AGTCCGATCATTGCAAATCTACC
PMM0710, pstS	GTACAGGTTGCAATGGGTATGG	TGTGAAAGCCTTAGTAGTTCCTG

Table 2.7: *Prochlorococcus* Med4 gene-specific RT-PCR primers.

followed by 40 cycles of 20 s of denaturation at 95°C, 30 s of annealing at 60°C and 45 s of elongation at 72°C.
For each sample, triplicate control PCRs were performed using 1 μl of a reverse transcription reaction without reverse transcriptase. To evaluate possible sample-to-sample variations in RNA concentrations, the concentration of the RNA subunit of ribonuclease P (*rnpB*) was used as an internal standard to normalise the mRNA levels.

Quantification
In a first step, the average of the CT values of the triplicate PCR reactions was calculated (mCT). If the expression of one gene is described relative to the expression of an internal standard, then the difference between the mCT of the investigated gene and the mCT of the internal standard was calculated: mCT(mRNA) - mCT(rnpB) = dCT. The equation 2^{-dCT} describes the initial cDNA amount which corresponds to the normalised amount of mRNA. To compare the transcription profiles of the same gene under different conditions (e.g. high light to normal light), one has to calculate the difference between the dCT value of the control, normal light, and the dCT value of the high-light culture. The transcript amount (cDNA amount) is then calculated from 2^{-ddCT}.

2.2 Bioinformatics part

2.2.1 Genome data

Cyanobacterial genomes were downloaded from NCBI GenBank (Wheeler et al., 2005) and for *Synechococcus* WH 7803 from Genoscope (Partensky, 2006).

2.2.2 Software tools

All used software tools are freely available. The multiple alignments were performed with CLUSTALW 1.81 (Thompson et al., 1994). The sequence logos were created with WebLogo (Crooks et al., 2006), described by Schneider and Stephens (1990); Crooks et al. (2004). All individual secondary structure predictions were done using MFOLD (Zuker, 2003). For visualisation of genomes and gene arrangements, the Artemis software (Berriman and Rutherford, 2003) appeared very useful.
All other scripts and tools implemented for more complex algorithms are described below and are available on request.

2.2.3 Promoter prediction

The whole genome prediction for putative promoter sites of marine cyanobacteria was based on the raster-score-filter method, described previously by Vogel et al. (2003a). A set of upstream regions (300 bp in length) was created using only non-coding sequences by masking the annotated coding regions. The information of experimentally determined and aligned -10 boxes was compressed in a scoring matrix. A search was done for promoters which exhibited features similar to the experimentally verified set, and the quality of the newly found -10 boxes inside uncharacterised regions was measured by means of this scoring matrix. The prediction was divided into three steps:

- raster, a search for 6 nt as a potential -10 region based on the known promoter structure in bacteria, -35 region - space - -10 region - space - TSS - space - translation start site

- score, determining the weights of the potential boxes with a scoring matrix

- filter, using a filter (cut-off value) to reduce the false positive rate of the predicted boxes

Scoring matrix
A set of binding sites can be used to create a scoring matrix with the nucleotides A, T, C and G as columns and the binding site positions as rows. Each entry of the matrix is determined by formula:

$$\log_2 \frac{p_i}{p_0} = \log_2 \frac{N_{observed}}{N_{random}} \qquad \text{[bits]}$$

with p_i as the observed frequency and p_0 as the expected frequency or $N_{observed}$ as the number of observed nucleotides and N_{random} as the number of expected nucleotides.
The determined score of one position within the binding site is the sum of the scores of one column of the score matrix and equals 2 bits minus the Shannon entropy H_1. Thus, if a position has a overall score of 2 bits, it is completely conserved; and zero bits stand for no conservation (equal distribution).
The entries of a score matrix can be extended by adding pseudocounts (small number of counts) to the columns of the matrix to increase the variability either to avoid zero counts or to add more variation than was found in the sequences used to produce the matrix (Mount, 2001):

$$\log_2 \frac{N_{observed} + N_{pseudocount}}{N_{random} + 4N_{pseudocount}} \qquad \text{[bits]}$$

It is a useful method, particularly suitable for small sets of known sites.

2.2.4 Phylogenetic footprinting

Phylogenetic footprinting is the major method for enriching for candidate regulatory elements by searching for conserved motifs upstream of orthologous genes from closely related

species (Bulyk, 2003). For *Prochlorococcus* Med4, MIT 9313, SS120 and *Synechococcus* WH 8102 a systematic intergenomic comparison was implemented to detect conserved sites between their orthologous upstream regions. The algorithm was developed in cooperation with Dr. Szymon Kielbasa using free software tools, such as BLAST (Altschul et al., 1990), GLAM (Frith et al., 2004) or MySQL (Axmark et al., 2006).

The phylogenetic footprinting algorithm can be divided in the following main steps:

- download of genbank files (Wheeler et al., 2005) and data management of annotated genomes by MySQL (Axmark et al., 2006)

- defining orthologous gene sets by pairwise BLASTp (Altschul et al., 1990) comparison (E-value $< 10^{-5}$), definition of reciprocal best hits and a single-linkage procedure (Tatusov et al., 1997, 2003)

- aligning the promoter sequences of orthologous genes using GLAM (Frith et al., 2004), a local alignment algorithm

- visualising and identifying segments of significant conservation with Graphviz (Ellson and North, 2006) as well as construction and comparison of position-specific weight matrices (PSWM) (Kielbasa et al., 2005)

- genome-wide searching for entries of candidate regulatory elements (Kielbasa et al., 2001) and definition of regulons

Palindromicity of all predicted sites was evaluated via PSWM comparison and chosen as a main criteria, as it is known to be correlated with the motif regulatory function (McGuire et al., 2000).

Additionally, all σ factors were identified by using The SEED database (Overbeek et al., 2005). Regulatory factors including transcriptional regulators were defined by comparison to the set of TIGRFAMs/Pfams (Fraser, 2006) in the functional role category "Regulatory functions" (Subroles: DNA interactions, Protein interactions, Small molecule interactions, Other).

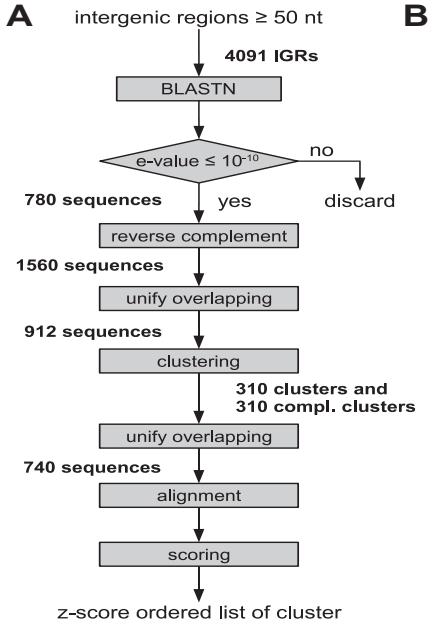
2.2.5 Prediction of ncRNAs

To identify candidates for experimental investigations, a comparative computational approach was developed in cooperation with Philip Kensche (Kensche, 2004; Axmann et al., 2005) that was based on sequence and structure conservation and used the program ALIFOLDz (Washietl and Hofacker, 2004). The genome sequences of *Prochlorococcus* SS120, Med4, MIT 9313 and *Synechococcus* WH 8102 were downloaded from NCBI GenBank (Wheeler et al., 2005). A summary of the computational screening is given in Figure 2.3.

A detailed description of the algorithm can be found by Kensche (2004); Axmann et al. (2005); the main steps are described below.

It was assumed that homologous RNA structures would show a reasonable degree of conservation on the sequence level for the given set of genomes. BLASTn version 2.2.8 (Altschul et al., 1990) was used to screen for local sequence conservations within intergenic spacer regions (IGRs) longer than 49 bp.
To take advantage of a multi-genome comparison, the pairwise sequence alignments were transformed into multi-sequence clusters via single-linkage clustering. Sequences that produced a significant blast hit (E-value $\leq 10^{-10}$) for a given query were collected into initial clusters.
Finally, each cluster was aligned using CLUSTALW version 1.81, default parameters (Thompson et al., 1994) and the resulting alignments were scored by ALIFOLDz (Washietl and Hofacker, 2004). The Z-score cut-off of -4 used by Washietl and Hofacker (2004) was chosen as a soft cut-off. For all structure computations, folding temperatures were set to 24°C, which is the approximate habitat temperature of the marine cyanobacteria studied here (Partensky et al., 1999).

Despite any structural conservation, any RNA in principle may encode for a peptide. The necessary reading frame as defined in this analysis consisted of at least ten consecutive codons starting with either one of the four possible start codons ATG, GTG, TTG or ATT and finishing with TAA, TAG or TGA. If a reading frame was present, the possible conservation of the encoded peptide sequence amongst other cyanobacteria was evaluated by alignments (tBLASTn). Only in the case of a conserved open reading frame, the RNA was considered to be coding and deleted from the list of possible ncRNAs.



B

	MIT	SS	WH
MED	337(30)	2179(250)	75(9)
MIT	-	189(26)	168(57)
SS	-	-	51(18)

Figure 2.3: Pipeline for comparative prediction of non-coding RNAs (Axmann et al., 2005). (a) Intergenic sequences (IGRs) longer than 49 base-pairs were gathered from four *Prochlorococcus* and *Synechococcus* genomes and locally aligned using BLASTn. Because of the initial asymmetric local alignment using BLASTn (see Figure 2.3b for a summary of significant BLASTn hits between the strains *Prochlorococcus* Med4 (MED), MIT 9313 (MIT), SS120 (SS) and *Synechococcus* WH 8102 (WH)), all candidate sequences were reverse-complemented. Redundancy in this data set was reduced by unifying those hits from each genome that showed a reciprocal overlap of 85 % or greater. This candidate set was used as both query and subject in another local alignment step (BLASTn considering only the query strand as possible subject strand). Sequences that directly produced a significant blast hit (E-value $\leq 10^{-10}$), or were connected by a chain of such hits, were gathered into clusters ('single-linkage clustering'). Both genome strands were screened; thus, the pipeline produced 310 pairs of clusters in both forward and reverse complementary orientation. After an additional unification step of overlapping sequences within each cluster, the resulting clusters and their complement clusters were scored using ALIFOLDz (Washietl and Hofacker, 2004). (b) The number of BLASTn high-scoring segment pairs for each query and subject combination of intergenic regions is given for a BLASTn E-value cut-off of 10^{-5} and after import of high-scoring segment pairs with an E-value of 10^{-10} or lower (in parentheses). MIT, *Prochlorococcus* strain MIT 9313; SS, *Prochlorococcus* strain SS120; WH, *Synechococcus* sp. WH 8102, MED, *Prochlorococcus* strain Med4 (Axmann et al., 2005).

Chapter 3

Results

3.1 Cyanobacterial promoters

The 5' RACE (rapid amplification of cDNA ends) is a very sensitive detection method for bacterial transcription initiation sites (TIS) by differentiating them from RNA processing sites (Bensing et al., 1996). This technique was successfully tested to identify TIS in *Prochlorococcus* Med4 (Vogel et al., 2003a) as the experimental basis for a computational promoter prediction. It was of special interest to check if this method could be transferred to another cyanobacterial strain and if different results would be obtained for distinct sets of genes, possibly linked to their divergent lifestyles.

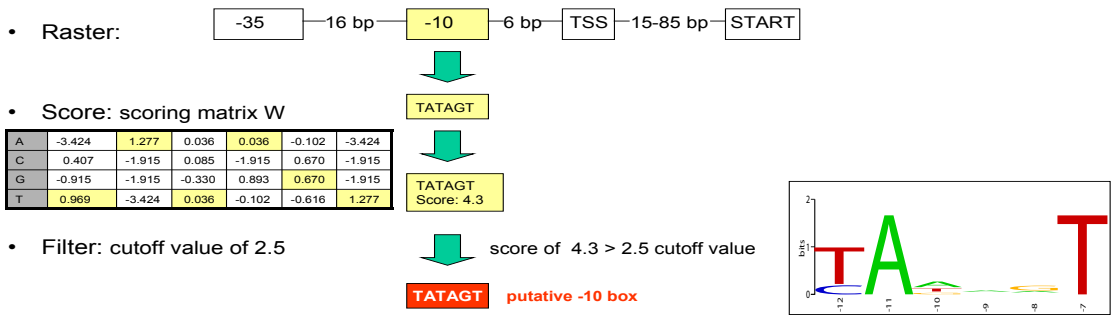


Figure 3.1: Raster-score-filter method (left) described by Vogel et al. (2003a) and weblogo of -10 boxes (right) of 10 promoter sequences of experimentally determined promoters in *Prochlorococcus* SS120. The positions -12 (T), -11 (A) and -7 (T) were found for the majority of TISs, see also Table 3.1.

The complete genome sequence of the extremely low-light-adapted *Prochlorococcus* SS120 was described in 2003 (Dufresne et al., 2003) including a manually verified genome annotation which could be used here to design gene specific primer-sets for RACE experiments in SS120. In parallel, an *in silico* prediction revealed a total of 3,130 transcription start sites for 1,289 noncoding upstream regions of the 1,930 predicted protein-coding or RNA genes in SS120 by using the raster-score-filter method (Vogel et al., 2003a), see also Figure 3.1,

with a slightly modified scoring matrix adapted to the AT-content of SS120 non-coding regions. The complete prediction can be downloaded from cyanolab (Hess, 2006). By analogy to Med4 (Vogel et al., 2003a), about 40 % of these sites were estimated to be functional.

The experimental validation by the RACE technique for eight randomly chosen genes in SS120 showed that promoter elements are very similar to the ones described for Med4 previously (Vogel et al., 2003a). Again a well conserved -10 region could be identified with a tANNNT like consensus motif, shown in Figure 3.1. As the computational algorithm was mainly based on this -10 region, the *in silico* prediction was correct for most of the -10 boxes upstream of the eight genes *pcbA, psbA, psbD, petH, kaiB, cpeB, ftsZ* and *ntcA* in SS120, listed in Table 3.1.

promoter	sequence	position
cpeB	gtaccatttctgctcttctagaagaggcattagagtTAGGTTggacagCG	-63/-64
ftsZ1	actttcttgaaaaacccaacagatctccaaagccagtCATAATgcaccAA	-31/-32
ftsZ2	cataatgcaccaaaaatcattcatctaaatcccgagTATGGTgatggctA	+7
kaiB1	acaagataacatcctcactcttagaggttaaatccgtCAAGATtagAtA	-36/-38
kaiB2	agaggttaaatccgtCAAGATtagAtAtatcaatTAAAGTctttctCtA	-14/-16
ntcA	gtcatttttgatacaagaggttttagcttcacttctTATCGTcctttgAA	-13/-14
pcbA	gtattgaaattagtccccttattttcgctcttttcgtTAGTGTttatgcG	-24
petH	tcataacgtgaactttgaaaaaggctatgtcgaataaTAAAGTtgagatA	-32
psbA	ttcttggttcagtacgcagttttgcaaactttcatcTAAACTcctcttGG	-55/-56
psbD	tgataaaaaatttccccgacttggttcttgaattgtTAAAGTctcaggcG	-30

Table 3.1: Experimentally determined mRNA 5' ends of *Prochlorococcus* SS120. Putative -10 regions are in upper case letters as well as the first transcribed nucleotide. The given positions are calculated relatively to the annotated start codon of the protein-coding gene.

In 2003 genomes were sequenced of another low-light-adapted strain, *Prochlorococcus* MIT 9313 (Rocap et al., 2003), as well as *Synechococcus* strains WH 8102 (Palenik et al., 2003), and recently WH 7803 (Partensky, 2006). All these strains were successfully analysed by RACE experiments, demonstrating the good practicability of the method for marine cyanobacteria. The huge amount of genomic and RACE data allowed a detailed comparative analysis of promoter regions for the different cyanobacterial strains. Subsequently, some of the most interesting results from this analysis are shown exemplarily.

3.1.1 *psbD* promoter region

Photosynthesis is the major energy source for an oxyphotoautotrophic bacterium. *Prochlorococcus* Med4 and SS120 possess the most reduced genomes of a free-living photosynthetic organism known to date. Therefore, nearly all photosystem II (*psb*) genes are single-copy, in contrast to other cyanobacterial genomes, which contain multiple copies of *psbD*, *psbA* and other genes (Dufresne et al., 2003).
By comparing the promoter regions of the different *psbD* (photosystem II D2 protein) genes (Fig. 3.2), the highest similarity could be detected between PMM1157, Pro1254,

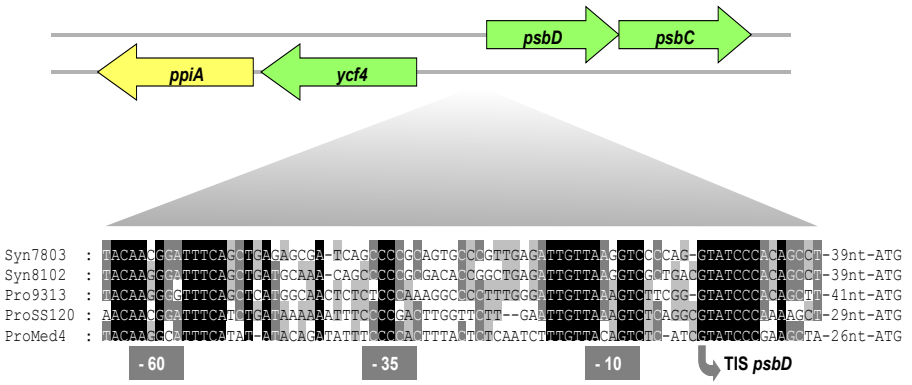


Figure 3.2: Arrangement of *ppiA, ycf4, psbD* and *psbC* and alignment of upstream regions of *psbD*. The green colour is indicating genes involved in photosynthesis. The arrangement of the four genes is conserved between five cyanobacterial strains. The alignment of upstream regions of *psbD* (photosystem II D2 protein) exhibits conserved blocks around TIS, -10, -35 and -60 region counted from the experimentally verified and aligned TIS (labelled by an arrow and described for *Prochlorococcus* Med4 by Vogel et al. (2003a); *Prochlorococcus* SS120 and MIT 9313, *Synechococcus* WH 8102 and WH 7803 this work). The total length of the intergenic spacer between *ycf4* and *psbD* varies around 200 bp within the five compared genomes.

PMT1179, SYNW0677 (WH 8102) and ORF1665 (WH 7803), although the strains WH 7803 and WH 8102 harbour an additional *psbD* gene with a verified and functional TIS. For all four most related copies, highly conserved regions occur at the TIS, the -10 and partly at the -35 region but interestingly also at the -60 region counted from the mapped and aligned TIS. An extended view of this genome-section (Fig. 3.2) revealed a very conserved microsynteny with a head to head arrangement of the *psbD* and *ycf4* (photosystem I assembly protein) gene with a distance of about 200 bp, which might represent a bidirectional promoter region and would allow a co-transcription of the genes needed for photosynthesis.
In agreement with these results, a reciprocal BLASTp search verified this set of *psbD* genes as the orthologous one. The additional *psbD* copies of *Synechococcus* do not possess the strong conservations within their promoter regions, which were found for the four orthologous regions. Therefore, these additional copies might be regulated in a different way.

3.1.2 *csoS1-rbcLS* promoter region

csoS1-rbcLS and *atpB* are further examples for highly conserved promoter regions and gene arrangements (Fig. 3.3 and Fig. 3.4). *rbcLS* encodes the ribulosebisphosphate carboxylase (large and small chain), the key enzyme of photoautotrophic carbon assimilation. The gene *csoS1* (or *ccmK*), and four additional genes (*csoS2-csoS3-orfA,B* for carboxysome peptides A and B) located immediately downstream of *rbcLS* encode components of the carbon concentrating mechanism. The alignment of five upstream regions of different strains (Fig. 3.3) reveals high conservations for the TIS, -10 and -35 elements, and again, several conserved positions at the -60 region.

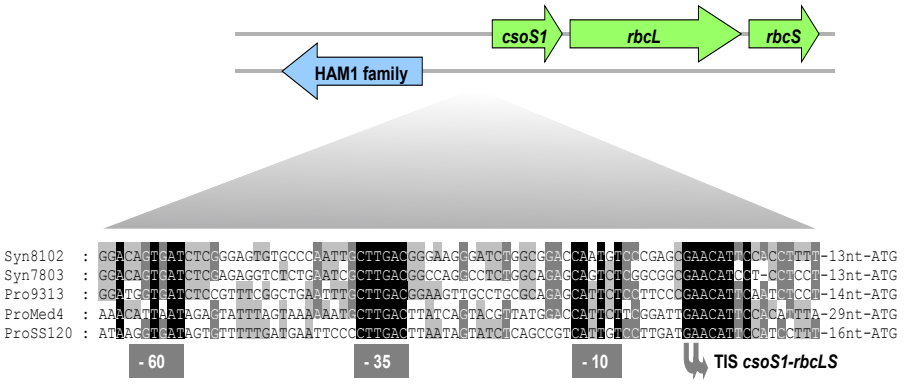


Figure 3.3: Comparison of the *csoS1-rbcLS* region from marine cyanobacteria. The *csoS1-rbcLS* operon is located in all cases head to head to a gene encoding a HAM1 family protein. The experimentally identified TIS (labelled by arrows) of *csoS1-rbcLS* for *Prochlorococcus* Med4 (Vogel et al., 2003a) and for *Synechococcus* WH 7803 (Garczarek et al., 2001) can be easily aligned with upstream sequences of *Prochlorococcus* SS120 and MIT 9313 and *Synechococcus* WH 8102 suggesting that the TIS, -10 and -35 region are conserved for these five strains. The total length of the intergenic spacer is 300 to 350 bp in the five compared genomes.

3.1.3 *atpB* promoter region

The arrangement of *groESL*, an operon encoding a cyanobacterial chaperone, on the forward strand and the genes *atpB, atpE* encoding the beta and epsilon subunits of the proton-translocating ATPase on the reverse complementary strand strand is the same for the five strains investigated. The length of the intergenic region between *groES* and *atpB* is about 220 bp, so that a co-transcription and therefore co-regulation might occur.

Again, within the upstream regions highly conserved positions can be found around the TIS and the -10 box. However, in this example nearly randomly distributed nucleotides appear at the -35 or at the otherwise often conserved -60 region, whereas a large 19 base-pairs long sequence block occurs between the -35 and -60 site.
Comparison of the predicted translation products of cyanobacterial *groES* and *atpB* demonstrated a very high amino acid identity with cognate chaperonins and ATPases (beta subunit), respectively, from bacteria and chloroplasts; their phylogenies confirmed the endosymbiotic origin of chloroplasts from cyanobacteria (Curtis, 1987; Webb et al., 1990; Morden et al., 1992).

3.1.4 *sfsA* downstream region

In contrast, for other examples non-coding sequences were observed without any significant degree of conservation despite a high degree of microsynteny of the respective genes, e.g. the non-conserved sequences downstream *sfsA* (Fig. 3.5), where a conserved arrangement with *mviN*, *amt1* and *lytB* was found. SfsA is similar to a sugar/maltose fermentation stimulation protein; *mviN* encodes a predicted membrane protein homologous to the virulence factor MviN; Amt1 belongs to the ammonium transporter family; and LytB is

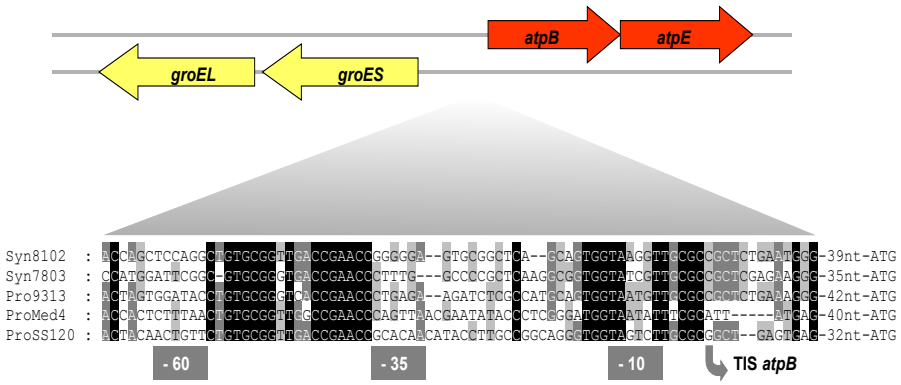


Figure 3.4: Arrangement of genes around *atpB* and alignment of *atpB* upstream sequences. This region exhibits a conserved synteny in *Prochlorococcus* Med4, SS120 and MIT 9313, *Synechococcus* WH 82012 and WH 7803 with a head to head orientation of *groES* and *atpB*. By aligning the 5' UTRs of *atpB* conserved boxes appear around the TIS and -10 region and between -35 and -60 region. The total length of the intergenic region between *groES* and *atpB* is about 220 bp.

described as a penicillin tolerance protein.

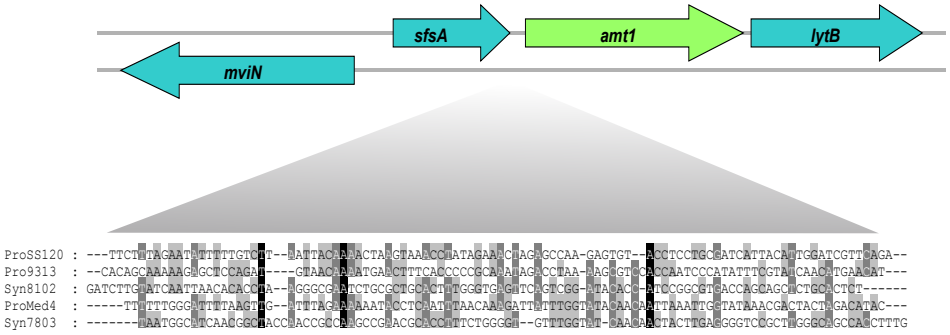


Figure 3.5: Synteny of *mviN*, *sfsA*, *amt1* and *lytB* and alignment of *sfsA* downstream sequences for five marine strains. The adjacent non-coding sequences downstream *sfsA* of *Prochlorococcus* Med4, SS120 and MIT 9313, *Synechococcus* WH 82012 and WH 7803 are shown. A conserved synteny with a head to head orientation of *mviN* and *sfsA* followed by *amt1* and *lytB* is visible. By aligning the downstream sequences of *sfsA* no conserved regions are detectable. The total distance between *sfsA* and *amt1* varies highly from about 100 bp (WH 8102) to 300 bp (SS120).

Thus, one can expect that not all comparable non-coding sequences of these five marine strains align and contain conserved blocks just by chance. Therefore, the phylogenetic distance might be assumed as far enough between the related non-coding regions given by this set of marine cyanobacteria, so that the detection of conserved regions as meaningful biological signals inside otherwise non-conserved sequence parts becomes possible.
A highly conserved promoter region might indicate a likely similar transcriptional regulation for the corresponding orthologous genes. On the other hand, several non-conserved

but regulatory important strain-specific sites, developed during a particular adaptation process, will be excluded by a comparative approach.

It seems obvious that not only protein-coding information appears to be conserved between genomes but also the essential regions for promoter recognition and transcription initiation. Interestingly, the strong conservation of the promoter regions between the different strains included mainly the transcriptionally important areas such as TIS, -10 or -35 region. Other nucleotides obviously follow different evolutionary constraints as there exist non-conserved parts between the conserved blocks of important promoter sites. Thus, a promising idea was to search more deeply for conserved elements upstream of orthologous genes to identify essential sites for regulation like *cis* elements recognised by transcriptional regulators.

3.2 Phylogenetic footprinting

A systematic intergenomic comparison of upstream nucleotide sequences from sets of orthologous genes (phylogenetic footprinting) was implemented to identify conserved regions and regulatory motifs other than the core promoter elements. Therefore, free software tools as BLAST (Altschul et al., 1990) or GLAM (Frith et al., 2004) were used to identify groups of orthologous genes and to analyse their upstream sequences by a local alignment algorithm. All intermediate data were managed by MySQL database server and in cooperation with Dr. Szymon Kielbasa.

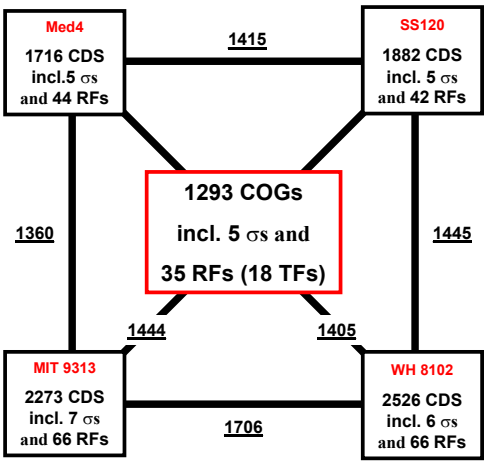


Figure 3.6: Overview of orthologous protein-coding genes including σ and regulatory factors, RFs, (including transcriptional factors, TFs) of four genomes, *Prochlorococcus* Med4, MIT 9313, SS120 and *Synechococcus* WH 8102, defined by best reciprocal BLASTp comparison with a cutoff E-value of 10^{-5}. The 4-way comparison resulted in 1293 clusters of orthologous groups (COGs) with at least one entry per genome (red box). Numbers of pairwise orthologs are underlined. In black boxes total numbers of CDS features of each single genome are displayed.

3.2.1 Orthologous gene sets

The starting point for a phylogenetic footprinting analysis is the definition of the set of orthologous protein-coding genes between the genomes of interest. Therefore, the reciprocal best BLASTp hits were identified based on the annotated CDS features of the four genomes *Prochlorococcus* Med4, MIT 9313, SS120 and *Synechococcus* WH 8102, similar to the system described by Tatusov et al. (1997, 2003). The 4-way BLASTp comparison with a cutoff E-value of 10^{-5} resulted in 1293 clusters of orthologous groups (COGs) with at least one entry per genome (Fig. 3.6). As a subset of these 1293 COGs, σ factors were identified by using the information of The SEED database (Overbeek et al., 2005). Regulatory factors (including transcriptional factors) were defined by comparison to the set of TIGRFAMs/Pfams (Fraser, 2006) in the functional role category "Regulatory functions". Thus, a core set of five σ and 35 regulatory factors including about 18 transcriptional factors was suggested as the regulatory potential mediated by proteins conserved between the four marine cyanobacteria (Appendix, Tab. 1).

3.2.2 Predicted TF binding sites and regulons

For each COG, its set of upstream sequences was extracted. The local alignment algorithm GLAM (Frith et al., 2004) detected the best conserved element between the four strains for each set of upstream regions. All GLAM-identified elements were translated into matrices and compared to each other. Thereby, clusters of similar matrices were observed, suggesting binding sites for *trans*-acting factors, which control more than a single gene. The Graphviz package (Ellson and North, 2006) appeared to be very efficient for the visualisation of calculated similarities between the matrices. Additionally, palindromicity of all predicted sites was evaluated, since it correlates frequently with a regulatory function (McGuire et al., 2000).
The best conserved elements present among different orthologous sets were extracted and realigned for construction of 21 position-specific weight matrices (PSWM); their weblogos are listed in the Appendix, Figure 2. Finally, these 21 scoring matrices were used to search for candidate regulatory elements in all upstream regions of all four genomes.
The results of the PSWM-search were analysed in detail by assigning the downstream genes to known pathways or regulons. Towards this goal, the genome annotation and an intensive literature search were informative. This final evaluation revealed three best motifs with analogy to already known sites for certain cyanobacteria: NtcA, LexA or ArsR, described in detail below.
These three predicted recognition sites similar to NtcA, LexA or ArsR are spaced and palindromic DNA motifs, and the further focus was kept on predicted elements, which possess these properties as well. Together with two additional best motifs their predicted binding sites and putative regulons are listed in Table 3.2. The complete list of results of the PSWM-search for the five best motifs is shown in the Appendix, Table 2.

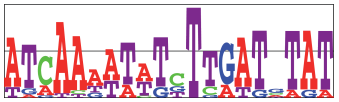
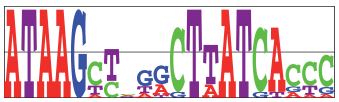
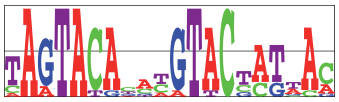
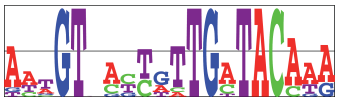
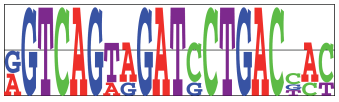
palindrome	sequence logo	best hits upstream of
ATCAA-N_6-TTGAT		*gap*; *arsR*; *pstS*; *phoB*; (Med4: *phoR*; SS120: *crp*)
ATAAG-N_5-CTTAT		*ham1*; *csoS1-rbcLS*; *petF*; *psbA*; (SS120: *hli12*)
TAGTACA-N_2-TGTACTA		*recA*; *umuD*; *umuC*; (SS120: *sbcDC*)
TGT-N_{10}-ACA		*ntcA; spt,agt; glnA; glnB; urtA*
GTCAG-N_6-CTGAC		conserved hypothetical proteins

Table 3.2: Top five of predicted motifs identified in *Prochlorococcus* Med4, SS120, MIT 9313 and *Synechococcus* WH 8102 via phylogenetic footprinting analysis. Only those hits are listed, which are highest-scoring in non-coding upstream regions and possess a predicted function for at least two orthologous downstream genes, indicating a putative regulon (hits found in one genome only are put in parentheses).

1. Motif: ArsR-like

Within the shared upstream region of *arsR* and *gap*, encoding a bacterial regulatory protein of the ArsR (arsenical resistance regulator) family and a glyceraldehyde-3-phosphate dehydrogenase respectively, a conserved inverted repeat was identified. The found motif ATCAA-N_6-TTGAT (Tab. 3.2) is identical to one of two direct repeats (spaced by 13 bp) of the ArsR site described for *Synechocystis* PCC 6803 by Lopez-Maury et al. (2003). The motif, identified here, often appears in two repeats but spaced by only two letters.
Interestingly, the PSWM-search predicted this site for several phosphate-depending genes: *phoB* (two-component response regulator, phosphate), *phoR* (two-component sensor histidine kinase, phosphate sensing), *pstS* (ABC-type phosphate transport system periplasmic component) and other genes encoding for putative transporters. A comparison to the DNA binding sites of PhoB, an eight-base motif repeat (TTAACCTT-N_3-TTAACCAT) suggested by Su et al. (2003) - similar to the Pho box of *Synechocystis* (Suzuki et al., 2004), revealed no similarity to the putative marine ArsR-like site, and moreover, PhoB binding sites predicted by Su et al. (2003) did not co-appear with it (with one exception: in Med4 upstream of *pstS*).
Arsenic, one of the most important global environmental pollutants, is widely distributed and also contaminating seawater. One of its chemical forms, arsenate, is a phosphate analogue, which might compete with phosphate in transport and in metabolic pathways (Thiel, 1988). Many organisms contain genes involved in arsenic resistance (*ars* genes). Thus, several cyanobacterial species are able to grow in the presence of high concentrations of arsenate and in low-millimolar concentrations of arsenite (Thiel, 1988; Lopez-Maury

et al., 2003). In *Synechocystis* PCC 6803 the *arsBHC* operon is responsible for arsenic sensing and resistance (Lopez-Maury et al., 2003). The complete *arsBHC* operon, regulated by ArsR in *Synechocystis*, is not present in the marine cyanobacteria analysed here, but two genes, *arsB* (arsenite transporter gene) and *arsC* (arsenate-reductase gene), exist separately within some of these marine genomes. The possible role of arsenic for marine cyanobacteria is still unknown, and no information is available about its uptake/export, toxicity, function and means of regulation there. Thus, a regulatory connection to the phosphorus-circuit remains highly hypothetical.
However, a complex transcription regulation was already observed during phosphate-starvation in other bacteria, which is presumably achieved by extensive cross-talk among multiple transcription factors (Birkey et al., 1998; Sun et al., 1996; Verhamme et al., 2002). For cyanobacteria, including *Synechococcus* WH 8102, a comparable complex network was suggested, in which the tandem DNA recognition by PhoB, a two-component signal transduction transcriptional activator, represents an important part of the P-regulon (Blanco et al., 2002; Suzuki et al., 2004; Su et al., 2003).
One might assume a repression of the phosphate-depending genes caused by ArsR binding in the promoter region similar to an effect observed in *Anabaena variabilis*, where a preincubation of phosphate-starved cells with arsenate caused subsequent inhibition of phosphate transport, suggesting that intracellular arsenate inhibited phosphate transport (Thiel, 1988).

2. Motif

The GLAM-detected motif for the spacer region of *ham1* and *csoS1-rbcLS*, ATAAG-N_5-CTTAT, was found further upstream of the -60 region, described for the *csoS1-rbcLS* promoter before. Additional entries were observed upstream of *petF* and *psbA*. Only in SS120 a site was predicted adjacent to *hli12*, a high-light inducible gene. Thus, this putative *cis* element might regulate genes needed for photosynthesis as well as *rbcLS*, the key enzyme of photoautotrophic carbon assimilation.
In *Synechococcus* sp. PCC7002 a CO_2 response element and the corresponding factor *trans*-acting on the *rbcLS* promoter were observed previously (Onizuka et al., 2002). Although the AT-rich element found there is not very similar to the here identified palindromic motif for marine strains, the negative regulatory role of its *trans*-acting factor, mediating the *rbcLS* transcription in response to CO_2 levels in the fresh-water strain PCC7002 (Onizuka et al., 2002), might characterise a mechanism existing in marine cyanobacteria as well.

3. Motif: LexA

The recognition site of another regulator, LexA, has been identified in recent years for several bacterial clades, demonstrating that the binding sequences of gram-positive and cyanobacteria are closely related (Mazon et al., 2004b). Further experimental and phylogenetic analyses revealed the existence of two independent evolutionary lines for the LexA recognition motif, suggesting gene loss and lateral gene transfer during the evolution of *lexA*, a gene governing such a complex regulatory network as the SOS system (Mazon

et al., 2004a). In *E. coli* most of the genes belonging to the SOS regulon are repressed by LexA. RecA, the positive regulator, activated by DNA damage, facilitates the autocatalytic cleavage of LexA, resulting in the expression of genes that are normally under the negative control of LexA (Walker, 1984).
LexA boxes were described previously for the cyanobacterial counterparts of the main SOS genes, *lexA, recA, uvrA* and *ssb*, in *Anabaena* PCC 7120 and *Nostoc punctiforme* ATCC 29133 (Mazon et al., 2004b). In contrast, for *Synechocystis* PCC 6803 only one LexA box was predicted by Mazon et al. (2004b), which is located upstream of the *lexA* gene. Moreover, the *Synechocystis* LexA protein contains modifications in important residues involved in the autocleavage process, which is especially important for LexA in its function as a repressor (Mazon et al., 2004b). Thus, based on detailed studies with *Synechocystis lexA, recA* and *ruvB* genes, Domain et al. (2004) argued against the involvement of LexA in the regulation of DNA repair in cyanobacteria. Recently, a complete novel function was reported for LexA, where it is not acting as a repressor but as an activator for the transcription of the *hox*-operon in *Synechocystis*, recognising two cyanobacterial LexA consensus sequences much further upstream of the core promoter (Gutekunst et al., 2005).
The LexA-binding sequence detected here for the marine strains is a palindromic and spaced motif with the improved consensus TAGTACA-N_2-TGTACTA as shown in Table 3.2. This consensus sequence is highly similar to the motif TAGTACT-AA-TGTTCTA, described for *Anabaena* PCC 7120 and located upstream *lexA* (Mazon et al., 2004b) as well as to the cyanobacterial LexA box discussed in detail under evolutionary aspects by Mazon et al. (2004a). Within the four *Prochlorococcus* and *Synechococcus* strains studied here, a putative LexA site was found upstream of the genes *recA* - encoding the bacterial DNA recombination protein, *umuC* and *umuD* - putative SOS mutagenesis proteins. Only a lower-scoring site was observed for the regulator encoding gene, *lexA* itself. Additionally, in SS120 this site was predicted upstream *sbcDC*, two genes for DNA repair exonuclease and ATPase, not conserved between the four marine strains.

4. Motif: NtcA

Another motif, GT-N_6TGNTACA, identified here by phylogenetic footprinting, was obviously similar to the palindromic consensus sequence TGT-N_{10}-ACA recognised by the transcriptional regulator NtcA. The motif predicted here for marine *Prochlorococcus* and *Synechococcus* includes additionally the flanking A/T-rich sequences and the conserved TG (or CA) dimer described previously (Vazquez-Bermudez et al., 2002). The best scoring hits of the genome-wide search for additional entries, using a PSWM for the 'marine' NtcA motif, are listed in Table 3.2: High-scoring sites were detected upstream of genes known to be regulated by NtcA or involved in nitrogen metabolism as *ntcA* - encoding the regulator itself, *spt,agt* - serine:pyruvate/alanine:glyoxylate aminotransferase, *glnA* - glutamine synthetase, *glnB* - nitrogen regulatory protein P-II and *urtABCDE* - putative urea ABC transport system.

5. Motif

The last of the five best clusters identified in this study possesses the conserved palindrome GTCAG-N_6-CTGAC. Most of the high-scoring sites for this motif are located upstream of genes with unknown function (conserved hypothetical proteins) indicating a new binding site for a new regulon in marine cyanobacteria, which is unknown so far and likely dissimilar from other bacteria.

3.2.3 Experimental verification of transcription initiation next to putative binding sites

5' RACE experiments in *Prochlorococcus* Med4 were used to locate the TIS of genes, for which putative binding sites had been predicted via phylogenetic footprinting. For two of the best scoring motifs, NtcA and LexA, three genes were chosen, respectively. The TIS of *urtA, glnA, ntcA* as well as for *lexA, umuD* and PMM1427, a conserved hypothetical ORF, were mapped by RACE experiments. The results are shown in Figure 3.7.

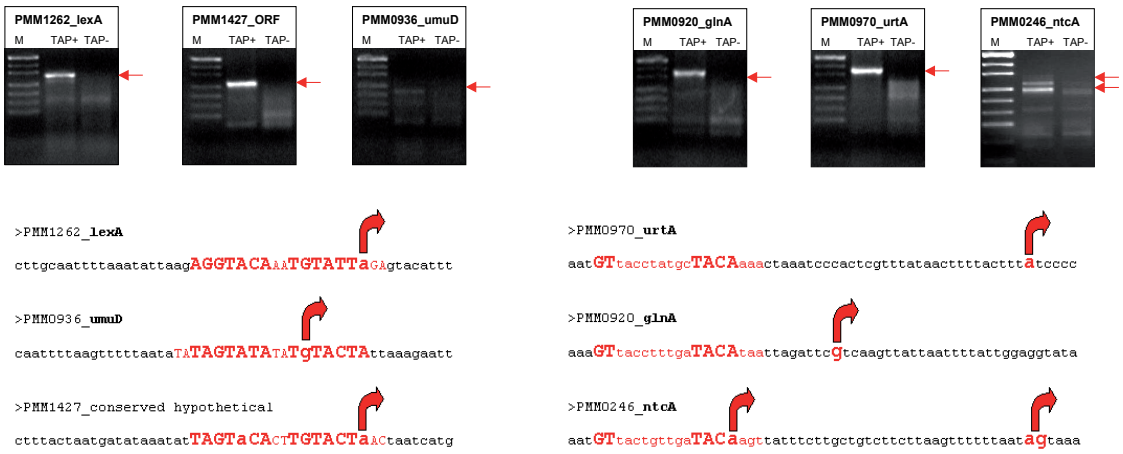


Figure 3.7: Results of the PCR step during 5' RACE experiments for *lexA, umuD*, PMM1427 (left panel) and *urtA, glnA, ntcA* (right panel) in *Prochlorococcus* Med4. For each gene one single TIS appears except for *ntcA*, which exhibits two signals in the TAP-treated (TAP+) line. Overlay of the putative LexA recognition sequence (upper case letters) and the determined TIS (indicated by an arrow) for *lexA, umuD* and PMM1427 (left panel). Overlay of the predicted NtcA binding site and the mapped TIS upstream of *urtA, glnA, ntcA* (right panel).

NtcA can act as both an activator and a repressor, as it is also the case for other regulators of the CAP family (Kolb et al., 1993; Herrero et al., 2001). In Med4, NtcA might act as a repressor if it binds to the -10 region, which could be the case for *glnA* and the second TIS (P2) for *ntcA* (Fig. 3.7). An activating function of NtcA is described by binding at the -35 region (Herrero et al., 2001), which might happen for *urtA* and the first TIS (P1) of *ntcA* (Fig. 3.7).
The putative negative control of *glnA* by NtcA in Med4 appeared unusual, because in

fresh-water cyanobacteria this gene was often found to possess a NtcA-activated promoter (Herrero et al., 2001). On the other hand, a second TIS might exist for *glnA*, which was not active under the conditions chosen for the RACE experiment. Additionally, an up-regulation of *glnA* expression was clearly demonstrated by RT-PCR for nitrogen-depleted Med4 cultures (by a factor of 5.8 compared to standard conditions), indicating a second NtcA-activated promoter site or a regulation different from those cyanobacteria investigated so far.
The *ntcA* gene itself has been found to be autoregulatory in several cyanobacteria. A negative feedback control, which is typical for bacteria, is suggested here for Med4 as well by the location of the NtcA recognition site at the -10 region of *ntcA* P2. If NtcA is activated upon nitrogen depletion, it is becoming competent in promoting the recognition of NtcA-dependent promoters (Herrero et al., 2001) and, therefore, its own expression increases (likely initiated at P1 in Med4). However, recent studies suggest an even more complex network for the nitrogen control in cyanobacteria, as NtcA seems to respond to a signal of the C-N balance of the cell rather than to the ammonium concentration (Herrero et al., 2001). Additionally, a cooperation with other factors might be possible. The group 2 sigma factor SigC of *Synechocystis* PCC 6803 was already shown to be involved in nitrogen-dependent promoter recognition (Asayama et al., 2004).

All three mapped TIS for *lexA, umuD* and PMM1427 are part of the predicted LexA binding site (Fig. 3.7). Thus, in case of LexA binding these genes would be repressed and might represent a part of the putative SOS regulon consequently.

3.3 Circadian rhythm and *kai* genes

The basic components described for the circadian clock of cyanobacteria are the clock proteins encoded by *kaiA, kaiB* and *kaiC*. The basic timing of circadian rhythm is mediated by KaiC phosphorylation as shown by *in vitro* experiments (Nakajima et al., 2005). The influence of transcriptional regulation is controversially discussed in the literature. Recently, the sufficient promoter region of *kaiBC* and the role of a constitutive negative element inside the *kaiA* coding region of *Synechococcus* PCC 7942 was described, suggesting these sequences to be essential for the circadian oscillation (Kutsuna et al., 2005).

A first step in this study was to identify these clock genes in the given set of marine cyanobacteria. Surprisingly, all *Prochlorococcus* strains sequenced so far do not possess a complete *kaiA* gene, although *kaiB* and *kaiC* are present at a similar genomic location as in the *Synechococcus* strains which possess all three clock genes (Fig. 3.8).

In all these cyanobacteria, the two- or three-gene *kai* operon is found on the reverse complementary strand upstream of the ribosomal protein operon starting with *rpl21*, coding for the 50S ribosomal protein L21. However, a detailed and comparative sequence analysis of the genomic location for the five marine cyanobacteria and the freshwater model strain for circadian clock, *Synechococcus* PCC 7942, clearly revealed a genome reduction during which the *kaiA* gene became deleted step by step: For the two *Synechococcus* strains, WH 8102 and 7803, the full-length *kaiA* gene is annotated, although the distance to the

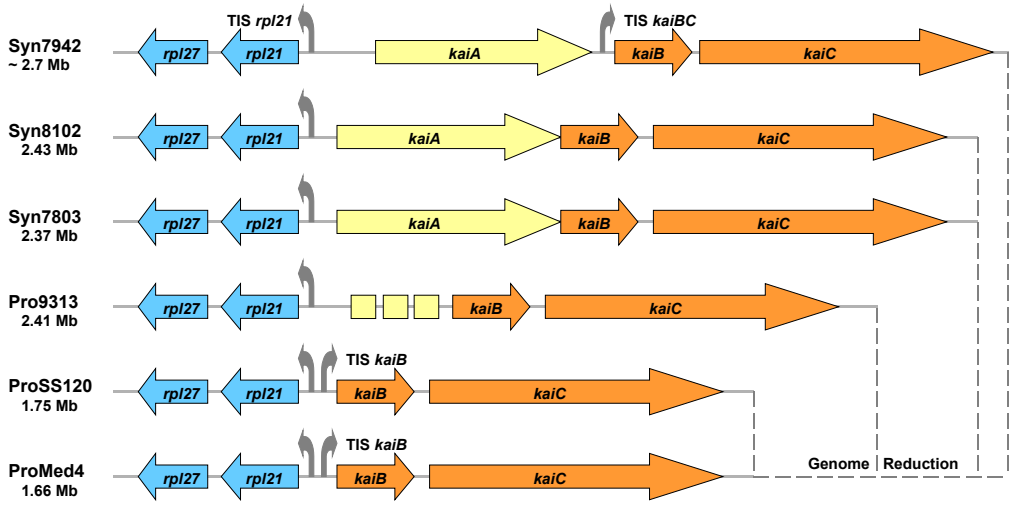


Figure 3.8: Arrangement of *rpl27* and *rpl21*, encoding 50S ribosomal proteins, and *kaiA*, *kaiB* and *kaiC* encoding the core clock proteins of cyanobacteria. From *Synechococcus* to *Prochlorococcus* a genome reduction is clearly visible in deleting the *kaiA* coding region with a complete gene loss in the smallest genomes, Med4 and SS120. The grey arrows indicate the identified promoter regions for *kaiBC* (Kutsuna et al., 2005), for *kaiB* and *rpl21*, this study, for Med4 (Vogel et al., 2003a) and for a predicted TIS of *rpl21* in PCC 7942.

adjacent *rpl21* gene is already reduced, compared to PCC 7942. For *Prochlorococcus* MIT 9313 only short relics of the *kaiA* coding region were detected (Fig. 3.9), and for *Prochlorococcus* SS120 and Med4 the complete coding region of *kaiA* is missing, resulting in a newly created intergenic spacer region between *kaiB* and *rpl21*. The reduction process in the *kai* gene operon, furthermore, correlates with decreasing genome sizes caused by an ongoing compactation starting from about 2.7 Mb for *Synechococcus* PCC 7942 to 1.7 Mb for *Prochlorococcus* Med4, the smallest known cyanobacteria (Fig. 3.8).

For the marine *Synechococcus*, no TIS for *kaiB* could be detected by RACE experiments. Also all attempts mapping TIS for *kaiA* or *kaiC* failed. Only secondary mRNA 5' ends resulting from RNA processing rather than initiation of transcription were found, indicating low expression and/or fast degradation of these mRNAs. The use of RNA samples from synchronised cultures collected at the maximum expression rate of the respective *kai* gene might improve the results of the RACE experiments.
In contrast, for the *Prochlorococcus* strains, SS120 and Med4, the transcription initiation sites of *kaiB* were mapped easily by 5' RACE. A well conserved motif could be detected next to the TIS of *kaiB* in Med4 and SS120 around position -60 of the promoter region, suggesting an important regulatory site of the *kaiB* promoter. Further analysis of this region identified this motif to be conserved between all six strains considered here (Fig. 3.10), although only Med4 and SS120 share a common promoter region for *kaiB* and *rpl21* (Fig. 3.8). Furthermore, promoter mapping for the adjacent *rpl21* gene revealed this site belonging to the 5' UTR of *rpl21*. The motif includes an inverted repeat, functionally it might correspond rather to the mRNA leader region of *rpl21* than constitute a part of the promoter for the *kai* operon. For instance, it might play an important regulatory role

for this ribosomal operon. Indeed, from enterobacteria it is known that functional boxes within the mRNA leader sequences mediate the autogenous control of ribosomal protein expression (Lindahl and Zengel, 1986; Zengel and Lindahl, 1994). Thus, a comparable regulatory mechanism for ribosomal genes may operate in marine cyanobacteria as well. In the case of Med4 and SS120 the promoter region of *rpl21* is teethed together with the promoter of *kaiB* as their core promoter elements overlay in part: TIS and -35 region overlap in Med4; -10 and -35 region in SS120. Thus, a combined expression of both genes may be assumed, which was already verified by RT-PCRs on RNA samples of synchronised Med4 cultures (Julia Holtzendorff, pers. communication).

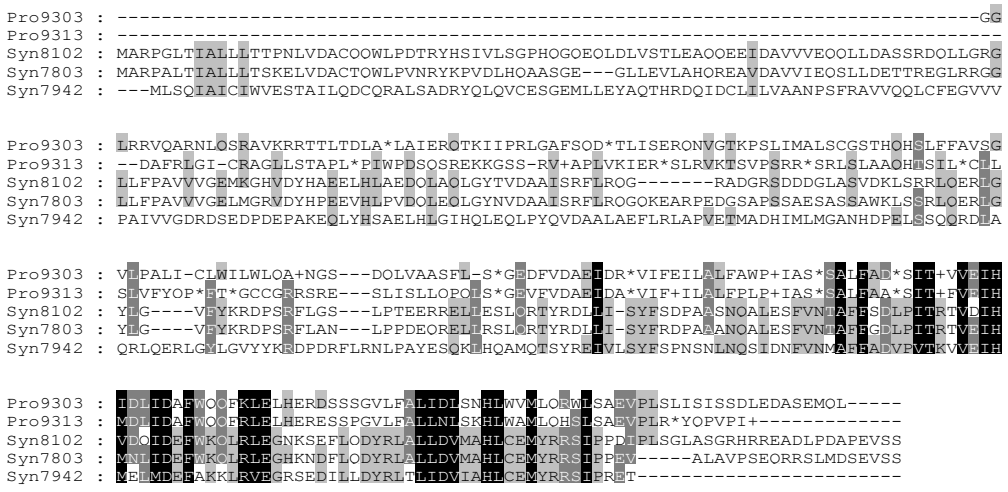


Figure 3.9: Alignment of the annotated coding regions of *kaiA* from *Synechococcus* PCC 7942, WH 8102 and 7803 together with the long spacer regions between the genes *kaiB* and *rpl21* of two low-light-adapted *Prochlorococcus* strains, MIT 9313 and 9303. A few amino acids conserved between all five strains are visible, whereby a putative reading frame for MIT 9313 and 9303 is interrupted by several stop codons (labelled by * and +), indicating the step by step deletion of the *kaiA* coding region in *Prochlorococcus* genomes.

Additional factors beside the Kai proteins of the cyanobacterial clock model are the circadian input kinases, CikA and predicted CikR, as well as the adaptive sensors, SasA and predicted SasR, which represent the key input and output pathway components, respectively. An important link to the metabolic status of the cell as well as a novel input pathway to the circadian oscillator is given by the light-dependent protein A (LpdA), recently described in *Synechococcus* PCC 7942 (Ivleva et al., 2005). A gene of the output pathway in the circadian system might be *cpmA*, whose mutation produced a modified circadian phasing phenotype for PCC 7942 (Katayama et al., 1999).

It was demonstrated earlier that strains of marine *Synechococcus* and *Prochlorococcus* exhibit a cell synchrony under dark/light conditions (Jacquet et al., 2001), although the existence of a real circadian clock has never been shown so far. Here, the complete genome

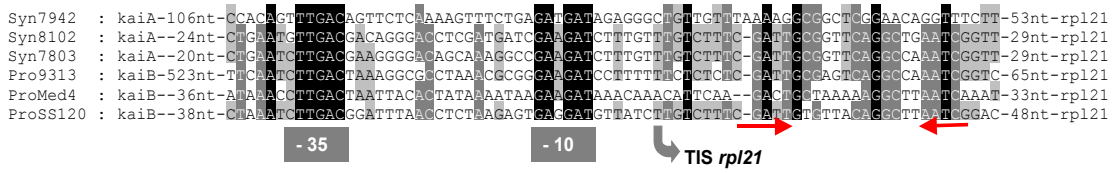


Figure 3.10: Alignment of the conserved *rpl21* upstream regions of *Synechococcus* and *Prochlorococcus*. Conservation of TIS (often thymine), -10 (GAaGAT) and -35 (cTTGAC) region is visible for all strains. These core promoter elements are highly similar to known *E. coli* standard promoters (except the thymine at TIS). Red arrows indicate a conserved inverted repeat within the *rpl21* 5' UTR, possibly indicating functional equivalence to the control elements of enterobacterial ribosomal operons (Lindahl and Zengel, 1986; Zengel and Lindahl, 1994).

sequences were used to screen for genes, that encode orthologs of other clock proteins or transmitting in- and output signals, described above. Among these periodosome components not only *kaiA* is absent in *Prochlorococcus* strains, but also the gene for the circadian input kinase, *cikA*, is missing in all marine strains analysed (Tab. 3.3). The genes *kaiB, kaiC, sasA* as well as *lpdA* and *cpmA* could be detected in all marine strains with reasonable e-values by BLASTp comparison to PCC 7942 coding sequences.

	PCC	7942	Med4	SS120	MIT 9313	WH 8102
core		*kaiA*	-	-	- (*)	SYNW0548
clock		*kaiB*	PMM1343	Pro1424	PMT1419	SYNW0549
genes		*kaiC*	PMM1342	Pro1423	PMT1418	SYNW0550
pathway	in	*cikA*	-	-	-	-
components	out	*sasA*	PMM1077	Pro1121	PMT1099	SYNW0753
signal	in	*lpdA*	PMM1560	Pro1714	PMT1730	SYNW2065
transduction	out	*cpmA*	PMM1278	Pro1352	PMT0358	SYNW1608

Table 3.3: Overview of orthologous genes in *Prochlorococcus* Med4, SS120, MIT 9313 and *Synechococcus* WH 8102 identified by BLASTp with known genes of *Synechococcus* PCC 7942, the circadian clock model. (*) In MIT 9313 only short relics of the *kaiA* coding region exist (Fig. 3.9).

The putative circadian clock of marine strains appears to be reduced by deletion of the key input pathway component, CikA, although another input pathway given by LpdA could compensate the absence of CikA. Additionally, in *Prochlorococcus* the coding region for the clock core component *kaiA* is deleted. Thus, a reduced periodosome might be found for a reduced and tiny genome of the slow-growing *Prochlorococcus* cell that divides on average once in 24 hours. The coupling of *kaiB* to ribosomal gene expression (*rpl21*) was suggested here, which may provide further evidence for a simplified circadian regulation mechanism in *Prochlorococcus*. Future experiments have to solve the question for an autonomous circadian oscillator existing in marine strains. The analysis of the protein components as well as studies on synchronised cultures could provide further insights into the marine pacemaker.

3.4 Transcriptional regulation of the cyanophage P-SSP7

The podovirus P-SSP7 is very host-specific in infecting the high-light-adapted *Prochlorococcus* strain Med4 exclusively. Recently, the genome sequence of P-SSP7 became available (Lindell et al., 2005) and allowed bioinformatic and experimental studies. Thereby, several similarities to the well studied coliphage T7 were observed (several T7-like class I, II and III genes including T7-like RNA polymerase), but also intriguing differences (host-like genes: *hli* and *psbA*; *int* gene and a possible integration site). Whether the transcriptional regulatory mechanism of T7 can be assigned to P-SSP7 is in dispute, because no significantly similar promoter sites were found for P-SSP7, suggesting that this phage does not belong to the otherwise highly conserved T7 group (Chen and Schneider, 2005). Thus, a detailed analysis of P-SSP7 promoter regions in comparison to those from T7 phages might give further insights into the transcriptional order of the P-SSP7 genome.

Each intergenic region ($>$ 50 bp) of coliphage T7 contains one or more transcriptional signals. All together, three strong early promoters and the early termination site for *E. coli* RNA polymerase as well as 17 promoters and one termination site for the T7 RNAP were studied intensively (Dunn and Studier, 1983). Thus, for cyanophage P-SSP7 transcriptional signals were assumed to exist in intergenic regions as well. Their prediction was based on a matrix (-10 box) for Med4 promoters (Vogel et al., 2003a), a matrix for T7-like promoters (Imburgio et al., 2000) and using the TransTerm algorithm (Ermolaeva et al., 2000), which detects rho-independent transcription terminators by searching for stem-loop-structures (inverted repeats) followed by a row of T's in the genome. In an experimental approach, the 5' ends of P-SSP7 RNAs were mapped using RACE (Bensing et al., 1996), a rapid amplification of cDNA ends, which allows a sensitive detection of transcription initiation sites and a differentiation from RNA processing sites.

The results are compiled in Table 3.4 including ten predicted bacterial promoter signals (-10 box), seven predicted termination sites (terminator) and six RNA 5' ends mapped experimentally (5' end). Surprisingly, no T7-like promoters were detected in P-SSP7 as well as no other frequently occurring sequence motifs, which is in agreement with findings of Chen and Schneider (2005), who suggested the transcriptional strategy for P-SSP7 to differ from the otherwise closely related T7 group. On the other hand, the DNA of P-SSP7 was shown to be transcribed from left to right during infection (Lindell, pers. communication) like T7 DNA, indicating a transcription of phage class I genes from bacterial promoter sites located near the left end of P-SSP7 DNA. Indeed, a Med4-like promoter was predicted near the left end, upstream of gene 1, although it could not be verified experimentally. This initial transcription might be terminated at predicted sites between gp12 and gene 3 or upstream of gp1, encoding RNAP, which is different in the T7 model, where the early expression of the phage RNAP is needed for the following stages of infection.

Non-coding gaps between coding sequences usually contain the transcription signals and also RNase III cleavage sites (Dunn and Studier, 1983). Thus, genes were chosen for

signal	DNA pattern	position	intergenic region
-10 box	cagaat	52..57	upstream gene 1
motif	ggttcaattcctctcccatcaattgc	91..116	upstream gene 1
5' end	t	113	upstream gene 1
motif	ggttcgactccctcccatccaattgc	1650..1675	gp12; gene 3
5' end	t	1672	gp12; gene 3
terminator	ccctgcaatgagcaggg	1675..1691	gp12; gene 3
-10 box	cacact	2591..2596	gene 6; gene 7
-10 box	caagtt	2629..2634	gene 6; gene 7
-10 box	taattt	3566..3571	gene 10; gp0.7
-10 box	tagagt	3577..3582	gene 10; gp0.7
-10 box	caatgt	3585..3590	gene 10; gp0.7
terminator	gcgattggtcaaaagccagtcgc	5006..5028	*int*; gp1
5' end	c	5032	*int*; gp1
-10 box	taattt	5034..5039	*int*; gp1
5' end	c	5045	*int*; gp1
5' end	c	5060	*int*; gp1
-10 box	tatgct	19430..19435	*psbA*; gene 28
motif	ggtgcgagtcctccctactcaatttg	19725..19750	gene 28; gp10
5' end	t	19749	gene 28; gp10
terminator	gggaccttcgggtccc	21031..21046	gp10; gp11
terminator	ggaagctgcattagatgtggcttcc	24626..24650	gp12; gene 32
terminator	gcacccgcaagggtgc	39402..39417	gene 49; gene 50
terminator	gcccttcagggc	41165..41176	gene 50; gp19
-10 box	tacttt	42978..42983	gp19; gene 52
-10 box	cactgt	42988..42993	gp19; gene 52
terminator	cccccgtaggggg	44063..44075	downstream *talC*

Table 3.4: Summary of transcriptional and other signals mapped to the cyanophage P-SSP7 genome. Signals for Med4 promoters (-10 box) were predicted based on a matrix (Vogel et al., 2003a), terminators by using the TransTerm software (Ermolaeva et al., 2000) and 5' ends by RACE experiments. The resulting DNA patterns are given as well as their absolute position in the genome and in the respective intergenic region. A conserved motif appeared by aligning sequences around detected 5' ends of gene 1, gene 3 and gene 29 (gp10). "gpX" are T7 designations for the homologous genes, and "gene X" designations correspond to the gene order for P-SSP7 as set out in Figure 1.5.

RACE experiments, which possess non-coding regions and/or predicted signals upstream, to identify 5' ends emerging from transcription initiation or processed sites. The resulting list contains 20 phage genes analysed experimentally:
gene 1, 3, 4, 8, 11 (gp0.7), 13 (gp1: RNAP), 14 (gp2.5), 17 (gp5), 19 (gp6), 20 (*nrd*), 24 (gp8), 26 (*hli*), 27 (*psbA*), 29 (gp10), 30 (gp11), 32 (gp13?), 36 (gp17), 50, 51 (gp19), 52 (designation given in Fig. 1.5).

RACE experiments for P-SSP7 revealed specific signals for gene 1, gene 3 (the same for gene 4) and gene 29 (gp10). These 5' ends (Fig. 3.11) might originate rather from RNase cleavage than transcription initiation as no specific differences were found whether TAP treatment was included or not. Furthermore, the mapped first nucleotide in each case turned out to be a uracil (or thymine in DNA) - a pyrimidine base, and not the purine usually found at an RNA primary 5' end resulting from initiation of transcription. An appropriate cyanobacterial -10 box similar to the tANNNT consensus sequence was absent in all three cases. Instead, an alignment of the sequences around these three sites resulted in a new motif, which is shown in Table 3.4 and discussed in detail below.

In T7 phages it was shown that the primary phage transcripts are processed at ten RNase III cleavage sites to provide the mRNAs observed *in vivo* (Dunn and Studier, 1983). RNA cleavage might be assumed for P-SSP7 as well, although a prediction of these sites is difficult, because a specific sequence pattern for RNase III is not known. On the other hand, signals predicted by TransTerm (Ermolaeva et al., 2000) include inverted repeats causing double-stranded regions, which might be recognised by RNase III as well.

3.4.1 Mapping of 5' ends by RACE experiments

Total RNA of *Prochlorococcus* Med4 and phage P-SSP7 was extracted at different time-points during lytic infection (Lindell et al., 2005) and kindly provided by Dr. Debbie Lindell as the basis for 5' RACE experiments. The primer sets for several phage genes (see above) as well as for a few host genes were chosen by using the annotated genome sequence of P-SSP7 (Sullivan et al., 2005) and Med4 (Rocap et al., 2003) to determine the 5' ends of their transcribed mRNAs.

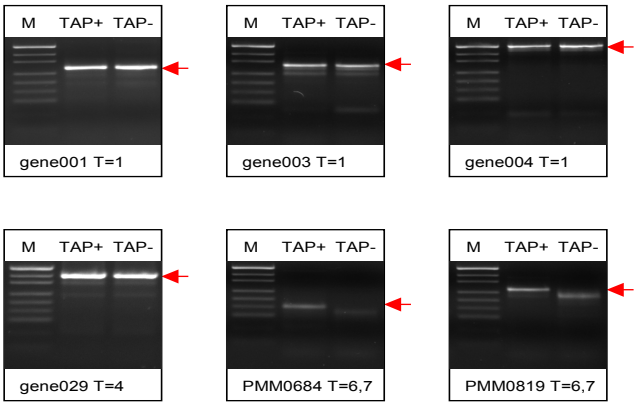


Figure 3.11: 5' RACE experiments for the cyanophage P-SSP7. Signals of maximum strength were detected for gene 1, gene 3, gene 4 and gene 29 (gp10) (designation given in Fig. 1.5) in the TAP-treated (TAP+) but also for the non-treated (TAP-) control sample, indicating processed but stable RNA 5' ends. The signals for 5' ends of the host (*Prochlorococcus* Med4) genes PMM0684 and PMM0819 are of lower strength and represent unprocessed RNA ends in the TAP+ variant (absent signal in the control variant, TAP-, for the same length).

Only for about the half of chosen genes RACE signals were observed (genes 1, 3, 4, 8, 13 (gp1), 17 (gp5), 20 (*nrd*), 29 (gp10)). Thereby, two different types of 5' ends may be distinguished:
The first group, upstream of gene 1, gene 3 (same 5' end as for gene 4) and gene 29 (gp10), see Figure 3.11, revealed signals of maximum strength by RACE experiments for TAP+ and TAP- (control) variants, indicating processed 5' ends with only one phosphate residue. Nevertheless, the three experimentally identified sites could be aligned to define a putative motif, shown in Figure 3.12b.

For all the other analysed sites (except for gene 13, gp1: RNAP; see 3.4.3, p. 62) no conserved region was identified, including those similar to known bacterial promoters or described phage promoters.

For the phage encoded *hli* and *psbA* genes, a signal could not be detected by RACE experiments, suggesting that they are either not transcribed alone, or their transcription is too weak for detection. Alternatively, they may be co-transcribed with the essential phage capsid genes. Indeed RT-PCR products spanning the intergenic regions between the phage *psbA* and capsid genes suggests that they are co-transcribed (Lindell et al., 2005).

Experiments using *Prochlorococcus* Med4 DNA arrays identified host genes up-regulated during infection (Lindell, unpublished): Are these Med4 genes transcribed by the phage RNAP or the host RNAP? The analysis of the 5' ends of host genes resulted in two perfect bacterial promoters for PMM0684 and PMM0819 (Fig. 3.11 and Fig. 3.12). Thus, a transcription by host RNAP is very likely: The -10 box fits perfectly to the previously derived model for Med4 promoter elements (Vogel et al., 2003a); the -35 region has a perfect match to the TTGAC sequence in the -35 region of *E. coli*, which also was the case for a subgroup of genes in Med4, like *rpl21* and *ccmK*. Interestingly, the upstream regions (Fig. 3.12) as well as the coding regions (not shown) of these two host genes are very similar to each other.
5' RACE experiments for three other Med4 genes, PMM0368, PMM1500 and PMM1501, gave no interpretable signal.

3.4.2 Putative motif upstream of genes 1, 3 and 29

The alignment of regions around the mapped 5' ends of gene 1, gene 3 and gene 29 (gp10) revealed a conserved motif, listed in Table 3.4 and shown in Figure 3.12. Interestingly, additional hits for the putative motif, which might also represent the possible RNAP binding site, do not exist within intergenic regions of the P-SSP7 genome, which was surprising as the related RNAP of phage T7 has 17 highly conserved promoters. On the other hand, the motif identified here is very uncommon: It has only weak similarity to the known T7 promoters, although the phage P-SSP7 is classified as T7-like and harbours the gene 13 (gp1) encoding a typical T7-like RNAP. The RNAP of T7-type phages is well-known to recognise highly conserved motifs. On the other hand, a computational search for more closely related sequence motifs failed to identify any candidates which might be more suitable as promoter sequences. Additionally, all three detected 5' ends of genes 1, 3 and 29 begin with an uracil base (thymine in DNA), which is the least likely case for all known transcription initiation sites.

It is already known that 5' ends for mRNAs of T7-like phage genes originate from multiple processes including bacterial and phage transcription as well as RNase III cleavage (Dunn and Studier, 1983). Especially, because the observed RACE signals for P-SSP7 gene 1, gene 3 and gene 29 (gp10) indicate RNA processing, it is likely that these 5' ends emerge from RNase cleavage. Inverted repeats are located next to the identified 5' ends (Fig. 3.12a). These are likely to form double-stranded regions, which might serve as recognition site for RNase III. One of five RNase III cleavage sites located in the early region of T7

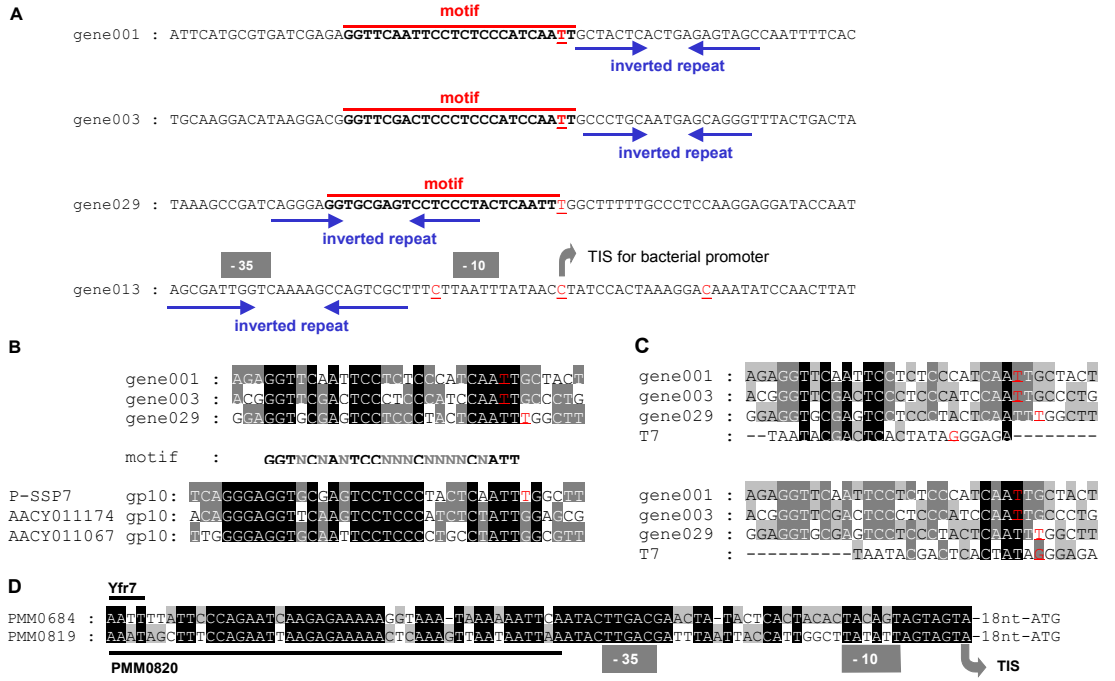


Figure 3.12: (A) 5' ends determined for the phage P-SSP7 gene 1, gene 3, gene 29 (gp10) and gene 13. Experimentally mapped start sites are red and underlined, predicted termination sites (inverted repeats) are marked by blue lines. (B) A consensus sequence is suggested for the upstream region of P-SSP7 gene 1, gene 3, gene 29 (gp10) (top). Motif searches inside a set of environmental samples (Venter et al., 2004) revealed two hits similar to the motif identified for the gene 29 (gp10) upstream region in P-SSP7 (bottom). (C) Comparison of the conserved sites upstream of P-SSP7 genes 1, 3 and 29 (gp10) to the consensus sequences of T7 promoter (Imburgio et al., 2000) revealed a very low similarity. (D) Upstream the two host genes PMM684 and PMM0819 of Med4 a perfect bacterial promoter site was detected including conserved -10 and -35 region. Adjacent coding sequences, gene for Yfr7 RNA and PMM0820, are indicated by black lines above and below the sequences.

was shown to possess a pU at the 5' terminus (Robertson et al., 1977), like it was found for 5' ends of P-SSP7 genes 1, 3 and 29. Additionally, the site for gene 1 was observed at each time-point during infection investigated, also at the early stage (T = 1, Fig. 3.11) at which the phage RNAP may not yet exist. If this would be the case, that phage RNAP is not transcribed early, then the promoter of gene 1 must be recognised by the host polymerase.

Thus, the motif might share more similarities to a RNase III cleavage site than to a T7 RNAP promoter site, and a verification of the motif origin in P-SSP7 could be highly advantageous. However, experiments like *in vitro* studies about P-SSP7 RNAP and its promoters have not yet been done and would be extremely challenging for this new phage-host system.

Therefore, further computational searches were attempted by using a set of environmental sequences from metagenomic analysis of the Sargasso Sea (Venter et al., 2004). An initial BLASTp search was done to identify orthologous genes of P-SSP7 gene 29 (gp10, encoding a capsid protein), which harbours the motif upstream of its coding region. For the four

newly found environmental gp10 homologs the upstream regions were aligned with the P-SSP7 sequence (upstream region of gene 29): Two of the environmental upstream sequences appeared to be very similar to P-SSP7 including the motif (Fig. 3.12b). Although these additional hits of the P-SSP7 motif can not be seen as a proof of its function, they may serve as strong evidence that this conserved site possesses a biological meaning for the phages, maybe for the transcription of the mRNA encoding the capsid protein, gp10.

By searching the Med4 host genome with the putative motif, not a single match was detected within non-coding regions, whereas three hits were found inside tRNA genes and tmRNA. One hit of the motif occurred at the 3' end of tRNA-Leu2, the tRNA which was suggested to serve as integration site for P-SSP7 (Sullivan et al., 2005). Another hit lies at the 3' end of tRNA-Gln1 and at the 3' end of the short tmRNA piece (mature two-piece tmRNA exists in *Prochlorococcus*, predicted by Gaudin et al. (2002) and shown in this work (Fig. 3.13 and Axmann et al. (2005)).
Both tRNA genes, tRNA-Leu2 and tRNA-Gln1, are each closely located to a downstream gene (*ndhL* and PMM1561, respectively), which might be co-transcribed with the tRNA, resulting in a mixed tRNA-mRNA polycistronic transcript. Thus, a later processing step at the tRNA 3' end to release the tRNA would be required. A similar situation was described for the *metY-nusA-infB* operon of *E. coli*, which is cleaved by RNase III in a hairpin structure downstream from the tRNA, separating the tRNA from the mRNA (Regnier and Grunberg-Manago, 1989).
The tmRNA of *Prochlorococcus* is encoded by one circularly permuted gene (Gaudin et al., 2002) and is found in an unusual two-piece form in the cells (this work, Fig. 3.13 and Axmann et al. (2005)). Thus, for the mature two pieces RNA processing is needed as well.
Taking all entries of the motif into account, one might suggest that the motif first observed in P-SSP7 seems to possess a biological function not only in the phage but also in its host, Med4, by representing a potential cleavage site for an RNase. In addition, one has to conclude that the motif identified in this study is also the closest candidate for a cyanophage-type promoter up the now.

3.4.3 Transcription of phage-specific RNA polymerase

Multiple 5' ends of phage mRNAs can be explained by a combination of processes, including bacterial and phage transcription and RNase III cleavage as known from T7 (Dunn and Studier, 1983). For P-SSP7, upstream of gene 13 (RNAP) three different ends were detected by RACE experiments (Fig. 3.12).
The region around one of these 5' ends appeared to be similar to a bacterial promoter site, which might represent a small hint for one of the most interesting questions about this host-phage system: Is the phage RNAP transcribed and translated prior to the other phage genes? The putative bacterial promoter upstream of gene 13 might indicate the possibility that the phage RNAP could be transcribed initially by the host RNAP and, moreover, independently of the other phage genes.
Another 5' end found for gene 13 is located immediately downstream of a predicted termination site. This termination site consisting of an inverted repeat may also serve as secondary structure for an RNase cleavage as it is described for several closely located promoter and cleavage site structures in the coliphage T7 (Dunn and Studier, 1983), see

also Figure 1.6b.

3.4.4 Missing promoter site upstream gene 30

The fact that the RACE failed in detecting a 5' end upstream gp11 (tail tubular protein) in P-SSP7 can be explained by comparison to the phage T7. A transcription termination site exists downstream gp10 (capsid protein) in T7 but no promoter for RNAP between this termination site and gp11 and gp12, so that these genes have to be transcribed by read-through of the termination site, which represents a part of the transcriptional strategy in T7-like phages. In P-SSP7 a termination site was also predicted downstream gp10 (Tab. 3.4), so that the absent promoter site for gene gp11 might resemble the T7 case.

3.5 Small RNAs in marine and other cyanobacteria

3.5.1 Known and abundant RNAs of marine strains

Total RNA samples from the four marine cyanobacteria *Prochlorococcus* Med4, MIT 9313, SS120 and *Synechococcus* WH 8102 were separated on high-resolution polyacrylamide gels to get an overview of the presence of small RNAs. This analysis showed abundant RNA molecules with sizes in the range 50 to 250 nt (Fig. 3.13). A particularly abundant class of RNAs in the 70 to 90 nt size range indicates the location of tRNAs in this gel, which was confirmed by hybridisation to the tRNASer [GCU]. The hybridisation signal for this tRNA was located at the upper end of this abundant cluster of bands, consistent with the fact that it is the largest annotated tRNA in these genomes. Several small RNAs migrated above the tRNA cluster and very few below it (indicated by the weakly visible bands below the tRNAs). These bands collectively indicated the occurrence of abundant small mRNAs, ncRNAs and precursors to tRNAs and rRNAs (Axmann et al., 2005).

Small eubacterial RNAs besides the abundant transfer and ribosomal RNAs, however, very rarely reach a concentration that allows direct identification in a gel. For known RNA species and their possible precursors or degradation products, information on their expression can be gained from hybridisation. Here, oligonucleotide probes were used for the scRNA and tmRNA and, as controls, the 5S rRNA and tRNASer [GCU], which was predicted to be the tRNA with the highest molecular mass. The lengths of the scRNAs in the four strains vary between 90 and 100 nt, in accordance with the varying lengths of the respective annotated *ffs* genes. The 5S rRNA was detected as a very abundant RNA species together with two precursors. Furthermore, the results of these Northern hybridisations confirmed that *Prochlorococcus* tmRNA is indeed composed of two separate molecules (Gaudin et al., 2002).

Several additional bands in the investigated size range indicated the presence of additional abundant small mRNAs or ncRNAs. The lack of specific oligonucleotide probes for hybridisation, however, made it difficult to get information about these. Thus, a computational prediction was used to identify candidates for further testing (Axmann et al.,

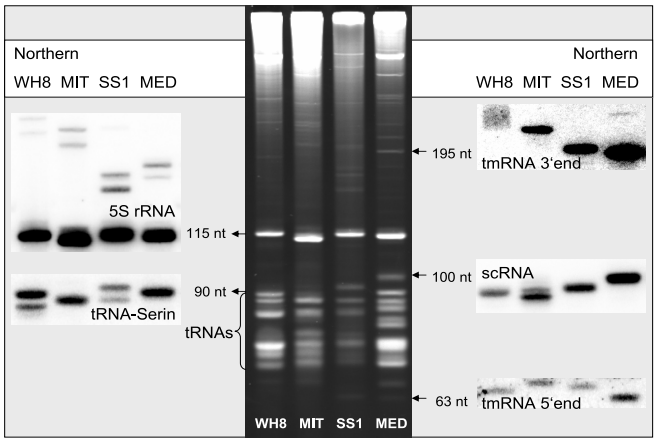


Figure 3.13: Small RNAs in marine Cyanobacteria, shown by Axmann et al. (2005). About 10 μg of total RNA from *Prochlorococcus* strains MIT 9313 (MIT), SS120 (SS1) and Med4 (MED) and from *Synechococcus* WH 8102 (WH8) was analysed by staining a 10 % polyacrylamide gel with ethidium bromide (center) and by Northern blot hybridisation with DNA-oligonucleotides directed against known RNA molecules such as scRNA (*ffs* gene product), the separate 5' and 3' ends of tmRNA and, as controls, tRNASerin and 5S rRNA. Two distinct precursors of the 5S rRNA were detected. Selected bands have been labelled by arrows in the hybridisation and in the gel picture and their sizes (nt, nucleotides) are indicated (Axmann et al., 2005).

2005).

3.5.2 Computational screening identified novel RNA species

The algorithm for identification of new, small RNAs was developed in cooperation with Philip Kensche (Kensche, 2004) and basically focused on sequence and structure similarities. Again, the advantage of a multi-genome comparison, between *Prochlorococcus* SS120, Med4, MIT 9313 and *Synechococcus* WH 8102, was used by transforming the initial, pairwise sequence alignments (BLASTn) into multi-sequence clusters via single-linkage clustering to increase the sensitivity of ALIFOLDz (Washietl and Hofacker, 2004; Washietl et al., 2005), the final scoring algorithm for each cluster. Also, single clusters were scored by ALIFOLDz, although the scoring method was found to be sensitive to the number of sequences in the alignment. Thus, a Z-score cut-off of -4 was considered as a soft cut-off for both, alignments and single clusters.
An overview of the computational screening is displayed in chapter Materials and Methods and a summary of the highest scoring clusters is given in the Appendix, Figure 3. Detailed information on all clusters predicted by this method, including the positions of all sequences, is available online (Kensche, 2005) and the terminology introduced there is followed here subsequently.

Although the sequence similarities between the predicted RNA elements in cyanobacteria and other organisms were weak, for many of the clusters, clues for their possible function could be obtained from the literature. These included elements that, according to loca-

tion or structure, might be functionally related to enterobacterial mRNA leader regions mediating the autogenous control of ribosomal protein and rRNA expression (clusters 5, 92, 227, 228) (Zengel and Lindahl, 1994; Lindahl and Zengel, 1986), the *rpoBC* leader region (cluster 245) (Barry et al., 1980) and the likely terminator (cluster 226). It was decided against direct experimental analysis of these elements, which are less likely to be novel types of ncRNAs. Additionally, two possible riboswitches for thiamine pyrophosphate (cluster 2) (Winkler et al., 2002) and cobalamin (cluster 101) (Nahvi et al., 2004) were excluded from further experimental investigations (Axmann et al., 2005).

3.5.3 Experimental verification of predicted ncRNAs in *Prochlorococcus* Med4

In the following, it was focussed on the analysis of small RNAs of the high-light strain *Prochlorococcus* Med4. In the remaining clusters, all candidate sequences from Med4 were tested by Northern hybridisation. Thereby, three different ncRNAs and a group of four related ones yielded strong signals with RNA preparations from Med4. Each of these seven candidate regions was probed for transcripts from both strands. Because some of these ncRNAs have a phylogenetic distribution beyond *Prochlorococcus* (see below), a new gene designation was introduced, *yfr* (for cyanobacterial functional RNA-coding gene), and Yfr for the respective RNAs. Each of these genes is discussed in detail in the following sections (Axmann et al., 2005).

Yfr1: a small RNA encoded between *guaB* and *trxA*

The *yfr1* gene was detected in three of the four cyanobacteria in the intergenic region separating *guaB* and *trxA* (Fig. 3.14). Although the two adjacent genes *guaB* and *trxA* are located in a similar genomic arrangement in SS120, a *yfr1* gene was not found at this or any other genomic position nor indicated by a Northern hybridisation signal. This result is in agreement with the high sequence divergence of the *guaB-trxA* intergenic spacer in SS120 compared to Med4, MIT 9313 and WH 8102 (Axmann et al., 2005).

The direction of *yfr1* is conserved between Med4, MIT 9313 and WH 8102. It is transcribed in the same direction as the mRNAs from two close-by neighbouring genes, indicating the possibility of co-transcription. Therefore, it was searched for the presence of specific transcriptional initiation sites (TIS) for *yfr1* and for *trxA* by rapid amplification of cDNA ends (RACE). A conserved TIS was mapped for *yfr1*, indicating that this transcript originates from a specific promoter (Fig. 3.14a) and reducing the likelihood that it is co-transcribed with *guaB*. Transcription of the adjacent *trxA* gene, encoding the redox regulator thioredoxin, was found to initiate approximately 100 bp downstream of the 3' end of the *yfr1* gene (Fig. 3.14a); co-transcription of *yfr1* with *trxA* is thus unlikely. In SS120, the lack of the *yfr1* TATA box, and the fact that the *trxA* TIS and TATA box are shifted upstream by about 20 bp compared to the other three strains (Fig. 3.14a), lends additional support for the absence of a *yfr1* gene (Axmann et al., 2005).

Strain	RNA gene name	Length of RNA in nt	Adjacent protein-coding genes	Orientation
MED4	yfr1	54	*trxA* and *guaB*	← ⇐ ←
	yfr2	94	PMM0363 and PMM0364	→ ⇒ →
	yfr3	95	PMM0686 and PMM0687	→ ⇒ ←
	yfr4	94	PMM0404 and *phdC*	← ⇒ →
	yfr5	89	PMM1027 and PMM1028	→ ⇐ →
	yfr6	244	PMM0659 and PMM0660	→ ⇐ ←
	yfr7	220	*purK* and PMM0684	→ ⇒ →
SS120	yfr2	90	*rpsU* and Pro0591	← ⇐ ←
	yfr6	239	Pro1007 and *purK*	→ ⇐ ←
	yfr7	175	Pro1007 and *purK*	→ ⇐ ←
MIT 9313	yfr1	57	*guaB* and *trxA*	→ ⇒ →
	yfr2	87	PMT1567 and PMT1568	→ ⇐ ←
	yfr7	175	*purK* and PMT0671	→ ⇒ ←
WH 8102	yfr1	56	*trxA* and *guaB*	← ⇐ ←
	yfr2	n.t.	overlapping SYNW1139	
	yfr3	n.t.	SYNW1140 and SYNW1141	→ ⇒ ←
	yfr7	174	SYNW1306 and *purK*	→ ⇐ ←

Table 3.5: Summary of identified ncRNA genes in *Prochlorococcus* MED4 and their orthologues in three related strains of marine cyanobacteria; n.t., not experimentally tested (Axmann et al., 2005).

Although direct information on cyanobacterial RNAs is scarce (Reyes et al., 1997; Golden et al., 1986) and not a single study exists for marine cyanobacteria, the half-lifes of eubacterial mRNAs are frequently in the range of a few minutes. In contrast, Yfr1 is extremely stable as a half-life of more than 60 minutes was measured after transcriptional arrest was induced by rifampicin (Fig. 3.15).
The expression of many bacterial regulatory RNAs is stimulated by varying environmental cues as by the stress response in which these RNAs have to play a role. Therefore, a variety of stress conditions and their possible impact on the accumulation of ncRNAs were tested.

Figure 3.16 shows a series of Northern hybridisations with RNA samples from cells that had been depleted of nitrogen, phosphate or iron, exposed to higher intensities of white or of blue light, or treated with 2 μM 3-(3,4-dichlorophenyl)-1, 1-N-N'-dimethylurea (DCMU) to induce oxidative stress or grown at elevated or lowered temperatures (30°C and 15°C). Normalisation of loaded RNA used 5S rRNA as an internal standard to compensate for small RNA sample loading differences; however, Yfr1 levels were unaffected by any of these conditions (Axmann et al., 2005).

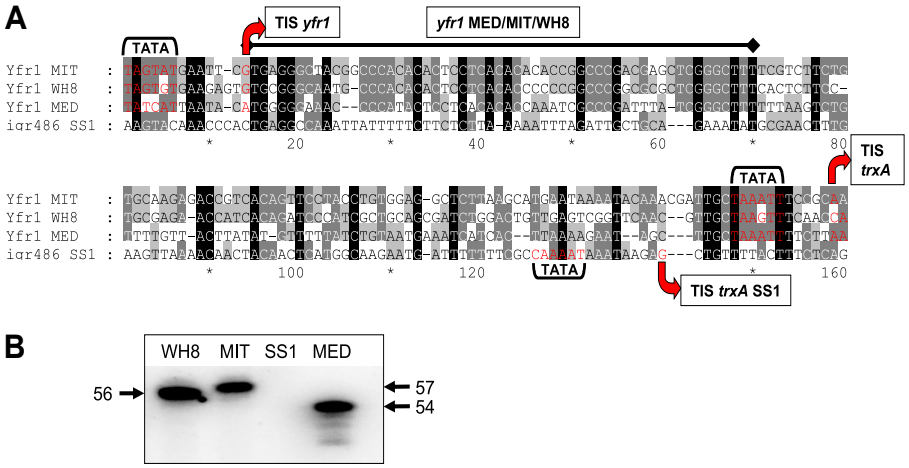


Figure 3.14: Experimental screen for the presence of an RNA-coding gene in the *guaB-trxA* intergenic region. (A) Sequence alignment of the *guaB-trxA* (*guaB*: sequence not shown, located upstream of *yfr1*) intergenic region visualises the conserved *yfr1* gene labelled by the bar above the alignment and its transcriptional initiation site in three of the analysed strains (MED, Med4; MIT, *Prochlorococcus* strain MIT 9313; WH8, *Synechococcus* WH 8102) but not in *Prochlorococcus* strain SS120 (SS1). Transcriptional initiation sites (TIS) and the deduced -10 elements are indicated. (B) Northern blots show a signal for Yfr1 at a size of 54, 56 and 57 nucleotides (nt) for Med4, WH 8102 and MIT 9313, respectively. No signal with RNA from SS120 confirms the absence of this gene in this strain, as was predicted from the sequence data.

Yfr2 to 5: A new family of related short RNAs

In top scoring cluster 194, a family of structurally highly similar RNAs (Yfr2, Yfr3 and Yfr4) was predicted (Fig. 3). Subsequent local alignments identified yet another similar sequence in Med4, and at least one homologue each in SS120, MIT 9313 and WH 8102 (Axmann et al., 2005).

Northern hybridisations with oligonucleotide probes specific for each of these candidate genes in Med4 yielded distinct bands of 89 to 95 nt. RACE mapping of 5' ends further confirmed that all four loci are transcribed in this organism (Fig. 3.17). The RNAs Yfr2 through Yfr5 in Med4 and their homologues in the other genomes are each encoded by distant genomic loci and the position of their genes is not fixed within the four investigated genomes with respect to adjacent genes (Tab. 3.5). The sequence comparison shows that Med4 Yfr2 and Yfr5 on one hand and Yfr3 and Yfr4 on the other hand are more similar to each other (Fig. 3.17a). The predicted secondary structures of the Yfr2-5 ncRNA family in MED4 are highly conserved with a GGAAACA repeat within the loop of the predicted 5' hairpin (Fig. 3.17c). Among the different tested environmental conditions, the amount of Yfr2-5 was affected by temperature (up at 15°C and down at 30°C) as well as by nitrogen limitation and incubation in blue light (Fig. 3.16) (Axmann et al., 2005).

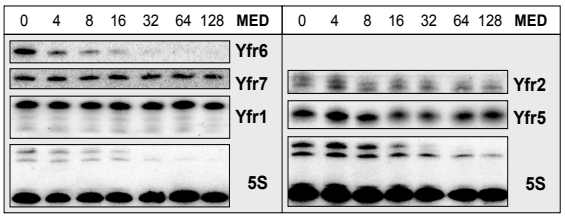


Figure 3.15: Determination of half lifes for Yfr 1, Yfr2 and Yfr5 to 7 in *Prochlorococcus* Med4, shown in Axmann et al. (2005). The time in minutes after the addition of rifampicin is given on top. Hybridisation to 5S rRNA served as control (Axmann et al., 2005).

Yfr6: A long RNA in Med4 and SS120

The *yfr6* gene was predicted in cluster 53 (Fig. 3). This cluster included nine different sequences, among which only *yfr6* in Med4 and SS120 may code for a functional RNA. The seven other sequences each have only about 40 nucleotide positions from their respective 5' untranslated region in common with Yfr6. That was sufficient to cluster all nine sequences together, but these other seven sequences included mRNAs for open reading frames. In contrast, Yfr6 from the two strains, Med4 and SS120, each have an extended sequence and structural similarity to each other (Axmann et al., 2005).

In Med4, *yfr6* is located between the hypothetical PMM0660 gene and PMM0659, the latter encoding 322 amino terminal residues of a DNA ligase. The region is framed by *trnS* and *nrdJ* (encoding a B12-dependent ribonucleotide reductase). In SS120, the *nrdJ-trnS* region lacks the *yfr6* gene, which instead is located 448 nt downstream of another ncRNA gene, *yfr7*. Despite the different genomic locations, Yfr6 sequences from the two strains show a nucleotide identity of approximately 70 % to each other. A Northern blot signal for Yfr6 is restricted to Med4 and SS120 and no signal was found in WH 8102 and MIT 9313. This 244 nt RNA had a half-life of approximately 2 minutes in Med4. In Med4, blue light and incubation in the cold elevated the expression of Yfr6 compared to white light or darkness. In addition, expression was reduced upon nitrogen depletion and under high light conditions (Fig. 3.16). The *yfr6* locus could also code for a 33 amino acid peptide as there is a possible reading frame that is conserved between Med4 and SS120 that begins at nucleotide 97 of the Yfr6 transcript in MED4. This situation, a relatively long transcript with strong structural potential and a very short centrally located reading frame, resembles the RNAIII from *Staphylococcus aureus*, a riboregulator from which the 26 amino acid delta-hemolysin peptide is also translated (Tegmark et al., 1998). In the hyperthermophilic archaeon *Sulfolobus solfataricus*, recently as many as 13 sense strand RNA sequences have been found that were encoded either within, or overlapping, annotated open reading frames (Zago et al., 2005).

Yfr7: A conserved RNA located downstream of *purK*

The *yfr7* gene is located downstream of *purK* (encoding phosphoribosylaminoimidazole carboxylase) in all four strains analysed here (Tab. 3.5). This conserved location of a ncRNA gene downstream of *purK* gene was already observed in freshwater cyanobacteria

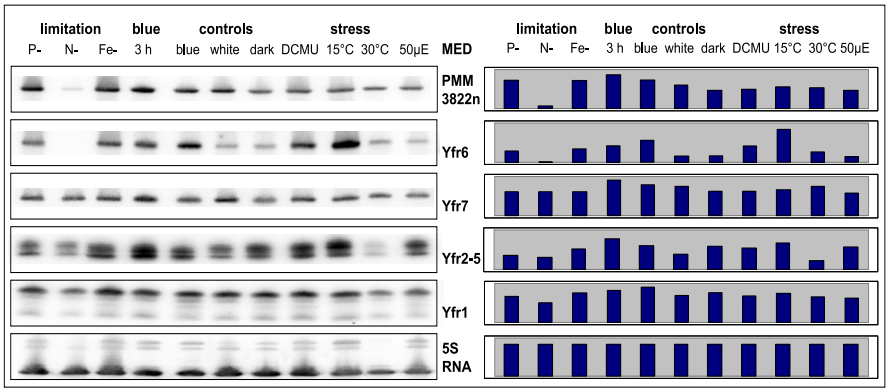


Figure 3.16: Test of transcript accumulation of Yfr1-7 from Med4 (MED) under different conditions, shown in Axmann et al. (2005). The left side shows the Northern hybridisations for which the following conditions were used: nutrient depletion (phosphate (P-), nitrogen (N-), iron (Fe-)); blue light for three hours (3 h); controls under blue (Blue), white (White) and no light (Dark); oxidative stress mediated by the application of DCMU; low (15°C) and high (30°C) temperatures; and high light intensity (50 μE). For comparison, 5S rRNA was hybridised as an internal standard and the mRNA of gene PMM3822n which, with a length of approximately 250 nucleotides, was taken as an example for a small mRNA. Additional controls by quantitative RT-PCR for the genes *isiB* (Fe), *glnA* (N), *pstS* (P) and *hli8* (high light) were carried out to confirm the effects of nutrient depletion or high light. The amounts of these mRNAs were enhanced by a factor of 79.7 (*isiB*), 5.8 (*glnA*), 2.8 (*hli8*) and 4.0 (*pstS*) under the respective treatment compared to standard conditions. Yfr6 shows an inconstant signal; for example, at cold, blue/white light, N-, Yfr2 to Yfr5 were hybridised with the consensus oligonucleotide y_gen (Fig. 3.17). The band intensities were quantified and normalised to the amount of 5S rRNA as an internal standard (right).

for the 6Sa RNA encoded by the *ssaA* gene (Watanabe et al., 1997).

At first, the search strategy identified this gene only in Med4 and SS120 (Fig. 3), due to the fact that in MIT 9313 and WH 8102 this corresponding region is located within annotated mRNA genes. These hypothetical genes, PMT0670 in MIT 9313 and SYNW1307 in WH 8102, are located on the forward strand. Their expression was not detected, but strong signals were found for Yfr7, which is transcribed from the complementary strand. The sequence of Yfr7 is highly conserved between the four strains (Fig. 4). Rifampicin tests showed this RNA to be stable (half-life $>$ 1 hour). In Med4, expression of Yfr7 was not affected by conditions employed in Figure 3.16.

3.5.4 Yfr1 exists in diverse cyanobacteria

Further computational screening detected additional entries of a *yfr1* gene in more distantly related cyanobacterial strains like *Synechococcus* PCC 7942, *Thermosynechococcus elongatus*, *Synechocystis* PCC 6803. Again, this small RNA was found to be encoded inside the *guaB-trxA* intergenic spacer. Compared to other eubacterial ncRNAs (Rivas et al., 2001; Tjaden et al., 2002), Yfr1 is one of the shortest bacterial ncRNAs, with a length of about 50 to 60 nt (Fig. 3.18b). No peptide reading frame within *yfr1* is conserved between any of the strains.

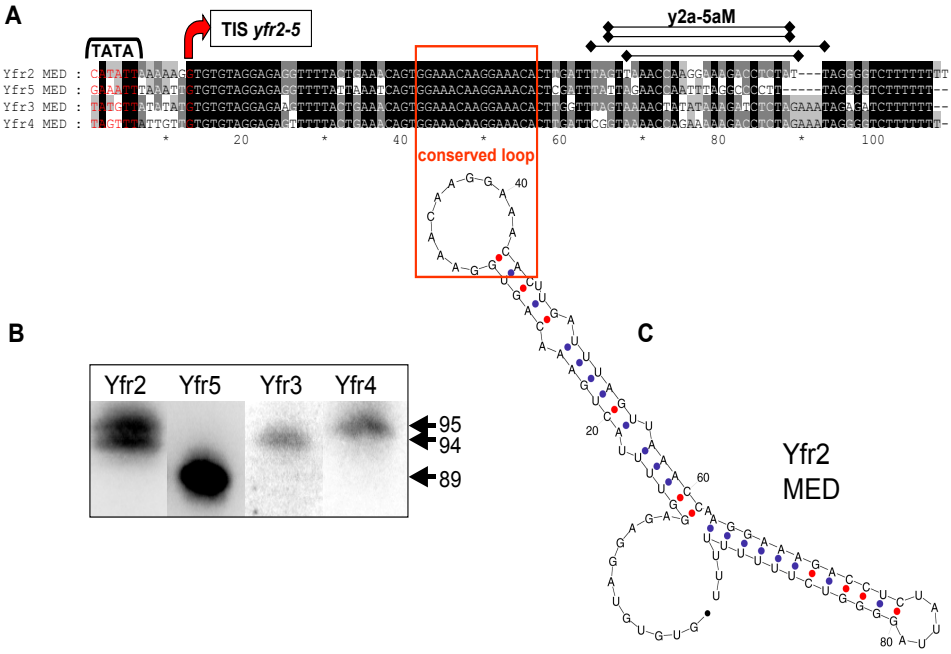


Figure 3.17: Comparison of Yfr2, Yfr3, Yfr4 and Yfr5 from Med4. (A) Sequence comparison of the *yfr2* through *yfr5* coding regions of MED4. Transcriptional initiation sites (TIS) and the deduced -10 elements are indicated. The location of specific oligonucleotide probes y2aM, y3aM, y4aM and y5aM used in Figure 3.17b and in 5' RACE and of the y_gen consensus probe used in Figure 3.16 is indicated by the lines with black diamonds on the ends on top of the alignment. (B) Signals for the four individual non-coding RNAs (ncRNAs) were detected in Northern blots using probes y2aM, y3aM, y4aM and y5aM. These probes have a minimum of five mismatches to their non-target ncRNAs, making cross-hybridisations impossible. The numbers indicate transcript lengths in nucleotides. (C) Prediction of secondary structure of MED4 Yfr2 by MFOLD (Zuker, 2003; Axmann et al., 2005).

Surprisingly, although Yfr1 sequences originate from diverse cyanobacterial strains, they share extensive structural and sequence conservation: Two terminal loops separated by a 16 to 19 nt unpaired region that contains a CACAC motif (Fig. 3.18). Consistently, the 3' located stem-loop element is formed by at least five GC pairs, and is followed by a short stretch of U residues, indicative of a Rho-independent transcription terminator (Fig. 3.18). Thus, a small but stable and structurally highly conserved RNA was identified here, which might exist in every cyanobacterial genome, with only one known exception: *Prochlorococcus* SS120.

3.5.5 Conservation of Yfr7/6Sa across the whole cyanobacterial lineage

The high sequence conservation of Yfr7 between *Prochlorococcus* Med4, MIT 9313, SS120 and *Synechococcus* WH 8102 was used to define oligonucleotides that hybridised to this RNA species in four additional, unsequenced strains of *Prochlorococcus* and in three additional *Synechococcus* strains (Fig. 3.19). The signal pattern is very distinct as all three *Prochlorococcus* strains, which are high-light-adapted ecotypes (Med4, MIT 9312, MIT

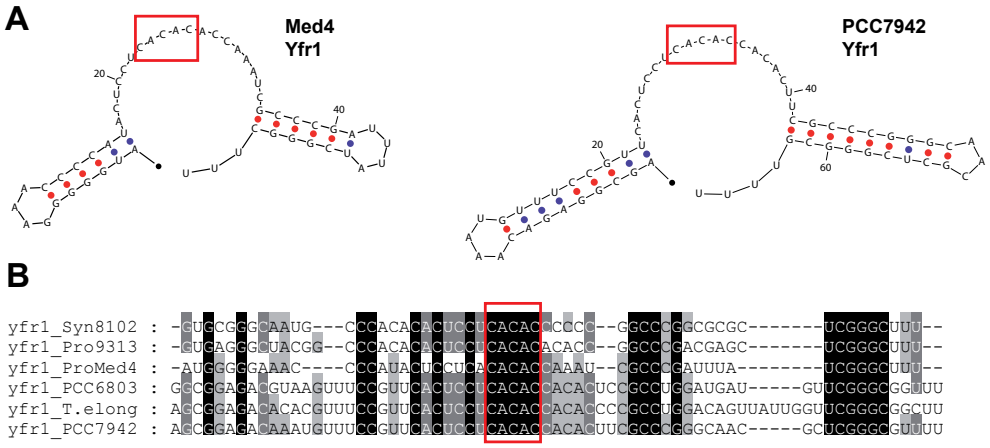


Figure 3.18: Identification of Yfr1 RNA in diverse cyanobacteria. (A) Two examples for the highly conserved secondary structure of Yfr1 predicted for *Prochlorococcus* Med4 and *Synechococcus* PCC 7942. (B) Parts of the *guaB-trxA* intergenic spacer were aligned and represent sequence similarities and a CACAC motif, which is located inside the unpaired region of Yfr1 (red box) of *Prochlorococcus* Med4, MIT 9313 and *Synechococcus* WH 8102 as well as in *Synechococcus* PCC 7942, *Thermosynechococcus elongatus*, *Synechocystis* PCC 6803.

9215), showed two signals in hybridisation, one at approximately 200 nt and one at approximately 300 nt, whereas RNA from the four low-light-adapted *Prochlorococcus* (SS120, MIT 9313, NATL2A and MIT 9211) and four *Synechococcus* (WH 8102, WH 7803, WH 8020, RS9906) strains gave a single signal at approximately 175 to 185 nt (Fig. 3.19). These strains represent a large genetic diversity within the marine cyanobacterial radiation (Fuller et al., 2003), thus the presence of real orthologues of Yfr7 in additional and even more distant cyanobacteria appeared likely.

Indeed, in the freshwater cyanobacteria *Synechococcus* PCC 6301 and *Synechocystis* PCC 6803, the 6Sa RNA has also been described directly downstream of *purK* (Watanabe et al., 1997). There is some structural similarity between Yfr7 and the 6Sa, which leads one to assume that these RNAs are real homologues of each other. In addition, a recent publication provided comparative structural information suggesting that the ncRNA Yfr7 described here and 6Sa RNA from the latter cyanobacteria have structural elements in common with the 6S RNA of γ-proteobacteria, in particular a large internal loop (the central bubble in Fig. 3.20c), a typical closing stem and terminal loop (Barrick et al., 2005). This possibly indicates that the here described 6Sa and Yfr7 RNAs are the orthologues of γ-proteobacterial 6S RNA and may have a similar role throughout the whole eubacterial radiation.

A detailed sequence comparison of already annotated 6Sa genes in very diverse cyanobacteria,

unicellular *Synechococcus* PCC 6301,

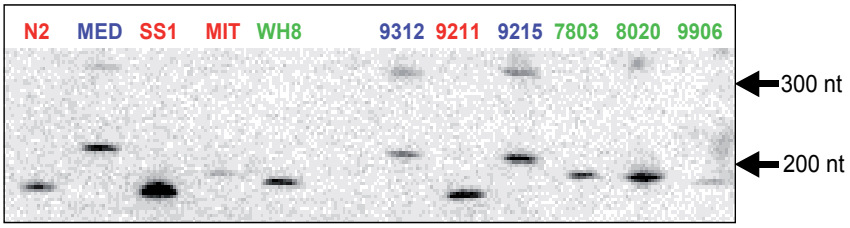


Figure 3.19: Identification of Yfr7 RNA in different marine cyanobacteria. In Northern blots, a signal for the predicted Yfr7 was detected with RNA from all four strains and seven additional strains from the marine cyanobacterial radiation: *Prochlorococcus* NATL2A (N2), Med4 (MED), SS120 (SS1), MIT 9313 (MIT), MIT 9312 (9312), MIT 9211 (9211), MIT 9215 (9215) and *Synechococcus* WH 8102 (WH8), WH 7803 (7803), WH 8020 (8020), RS9906 (9906). The high-light-adapted *Prochlorococcus* strains are labelled in red, low-light-adapted strains in blue, and *Synechococcus* strains are colour-coded in green.

unicellular, non-nitrogen-fixing *Synechocystis* PCC 6803,
unicellular, thermophilic *Thermosynechococcus elongatus*,
unicellular, unusual *Gloeobacter violaceus* PCC 7421 and
filamentous, heterocyst-forming, nitrogen-fixing *Anabaena* PCC 7120,

together with the described marine Yfr7 RNAs did not result in an alignment exhibiting well conserved regions, which were assumed for the 6Sa stem structure.

Further analysis of the genome location and alignments including the complement sequences (Fig. 3.21) indicated that the *ssaA* gene was annotated on the incorrect strand for *Thermosynechococcus elongatus*, *Anabaena* PCC 7120 and *Gloeobacter violaceus* PCC 7421; which was supported by Barrick et al. (2005) very recently.
Using the complement sequences of annotated *ssaA* genes for these three strains yielded a significantly improved alignment of 6Sa RNAs. This initial alignment was extended by adding BLASTn hits of 6Sa in *Synechococcus* PCC 7942 and *Nostoc punctiforme*. Additionally, sequences of the four analysed marine strains as well as their BLASTn hits in *Prochlorococcus* MIT 9312 and *Synechococcus* WH 7803 were added to the 6Sa alignment (Fig. 4), which was the basis for further phylogenetic studies as the Parsimony tree shown in Figure 3.21. The 6Sa tree clearly separates the chosen cyanobacteria into two clusters: The highest relatedness was found inside the marine group of strains and the freshwater/terrestrial group. Thereby, the location of the *ssaA* gene appeared to be conserved downstream of *purK* except for a few cases: *Synechocystis* PCC 6803, *Gloeobacter violaceus* PCC 7421 and *Thermosynechococcus elongatus*, which interestingly appear close to each other within the 6Sa tree (Fig. 3.21).

The highly conserved regions of the 6Sa/Yfr7 alignment were used to design oligonucleotides that hybridised strand-specific to 6Sa RNA in all sequenced and unsequenced strains available for culturing, whereas probes against the incorrect orientation failed. Thus, the expression of 6Sa RNA could be verified for all strains investigated by hybridi-

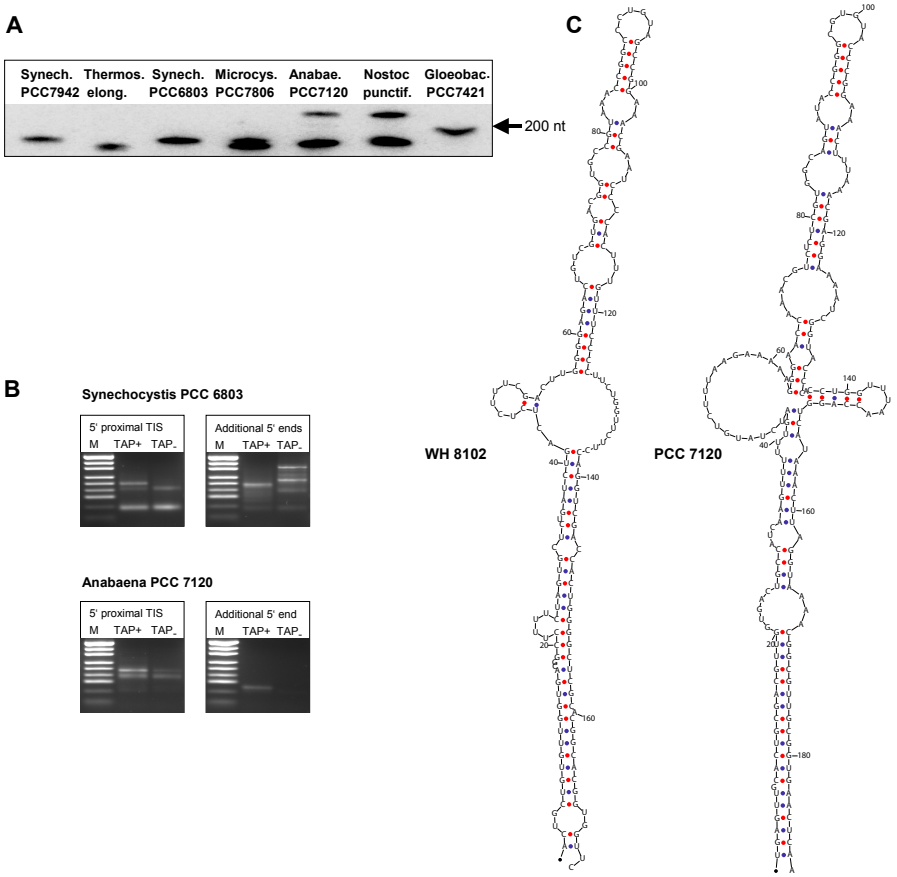


Figure 3.20: Identification of 6Sa RNA in different cyanobacteria. (A) Hybridisation of Northern blots with radiolabelled probes designed for a wide range of cyanobacteria identified signals for 6Sa in each strain investigated (from left to right: *Synechococcus* PCC 7942, *Thermosynechococcus elongatus*, *Synechocystis* PCC 6803, *Microcystis* PCC 7806, *Anabaena* PCC 7120, *Nostoc punctiforme* and *Gloeobacter violaceus* PCC 7421). (B) RACE experiments detected the TIS of 6Sa RNA in *Synechocystis* PCC 6803 and *Anabaena* PCC 7120 as well as additional processed 5' ends. (C) Prediction of secondary structure of the *Synechococcus* WH 8102 ncRNA Yfr7 and of the *Anabaena* PCC 7120 6Sa RNA.

sation to Northern blots, shown in Figure 3.20a. Thereby, the correct orientation of the *ssaA* gene, as described above, was experimentally proven for *Thermosynechococcus elongatus*, *Anabaena* PCC 7120 and *Gloeobacter violaceus* PCC 7421.

Additionally, RACE experiments verified the expression of 6Sa in *Synechocystis* PCC 6803 as annotated and in *Anabaena* PCC 7120 from the complement strand and detected the TIS for 6Sa as well as additional processed 5' ends (Fig. 3.20b). Their determined promoter regions upstream of *ssaA* harbour well conserved elements for the -10 (TATAAT) and -35 region (TTGac), perfectly matching to the consensus sequences of standard promoters in *E. coli* as well as a few promoters (like *rpl21* and *ccmK*) found in *Prochlorococcus* Med4 (Vogel et al., 2003a).

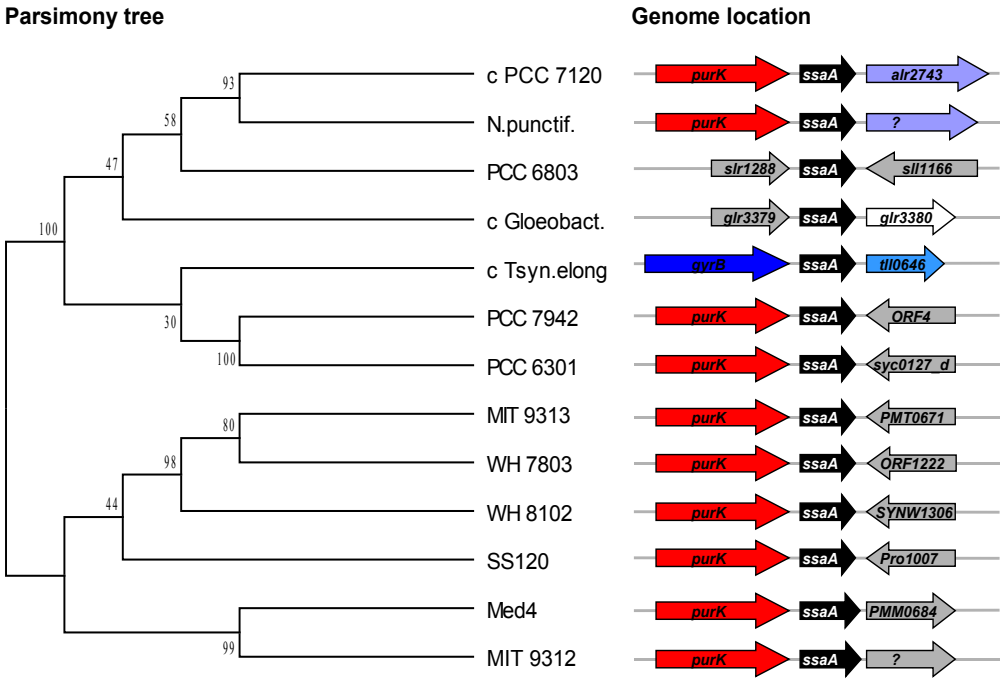


Figure 3.21: 6Sa/Yfr7 Parsimony tree is shown based on the alignment of 13 sequences from (top to bottom) *Anabaena* PCC 7120, *Nostoc punctiforme*, *Synechocystis* PCC 6803, *Gloeobacter violaceus* PCC 7421, *Thermosynechococcus elongatus*, *Synechococcus* PCC 7942, *Synechococcus* PCC 6301, *Prochlorococcus* MIT 9313, *Synechococcus* WH 7803, WH 8102, *Prochlorococcus* SS120, Med4 and MIT 9312. Additionally, the arrangement of the 6Sa/Yfr7 encoding *ssaA* gene and adjacent genes is illustrated for the 13 strains.

Day-time and growth-phase dependent expression of Yfr7 in *Prochlorococcus* Med4

RACE experiments of Med4 RNA samples detected two 5' ends for Yfr7 (Fig. 3.22a). At the proximal TIS, a 220 nt long RNA is transcribed, and further upstream inside the *purK* gene, a 5' end resulting from processing yields a 332 nt RNA (Fig. 3.22b).

A comparable situation was recently described for *Bacillus subtilis*, where two differentially expressed 6S RNAs exist, 6Sa and 6Sb (Barrick et al., 2005). Each transcript is encoded by its own gene, whereby 6Sb was identified as the orthologous RNA to *E. coli* 6S, whereas 6Sa appeared functionally diverged (Barrick et al., 2005). For both *Bacillus subtilis* transcripts a different growth-phase dependent expression was demonstrated indicating an additional fine tuning of the transcriptional response to nutrient limitation (Barrick et al., 2005).
On the other hand, experiments showed Yfr7 to be stable in Med4 and its expression was not affected by stress conditions including nutrient limitation (Fig. 3.16). Therefore, further investigations were needed to find out if expression of Yfr7 is affected by changing growth conditions as demonstrated for 6S in *E. coli* or *Bacillus subtilis*.

Thus, the expression of both Yfr7 transcripts was tested during growth of the Med4 strain

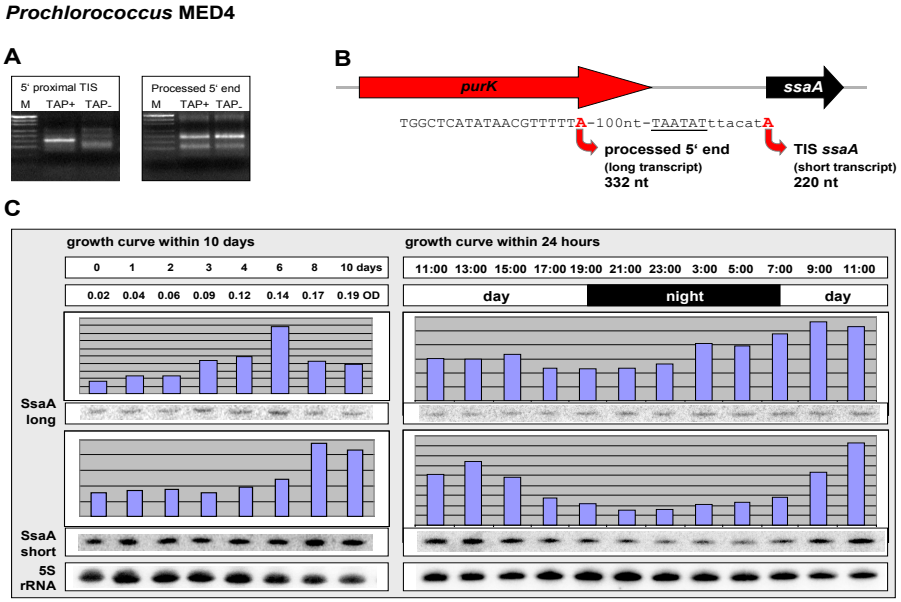


Figure 3.22: Identification and transcript accumulation of Yfr7 in *Prochlorococcus* Med4. (A) 5' RACE detected the TIS of the *yfr7* gene as well as the processed 5' end further upstream. (B) Genome location of *yfr7* downstream of *purK* is shown. Transcription starting at the proximal TIS resulted in a 220 nt short RNA. The long transcript variant processed at a site inside the *purK* gene has a length of 332 nt. (C) The accumulation of both Yfr7 transcripts during growth within 10 days (left panel) and within 24 hours (right panel) in a synchronised culture is shown. Northern blots were hybridised with radiolabelled probes specific for Yfr7 and 5S RNA. Band intensities were quantitated, corrected for 5S loading controls, and normalised to the maximum level of the respective Yfr7 transcript (%). The levels of both Yfr7 transcripts are different for each timepoint and are time-shifted. The 332 nt transcript peaks at OD_{600nm} 0.14; the 220 nt RNA accumulates at the last time-points, day 8 and 10 (left panel). Within a day (right panel) both transcripts are changing their amounts rhythmically with peaks in the morning for the 332 nt RNA and at noon for the shorter 220 nt variant.

(Fig. 3.22c). Interestingly, the growth curve within ten days from OD_{600nm} 0.002 to 0.19 revealed varying expression levels with different maxima for the long and the short Yfr7 transcript: The total levels of the processed 332 nt RNA are peaking at OD_{600nm} 0.14, probably the entry into the stationary phase under these growth conditions; at the latest time points (day 8 and 10) the 220 nt RNA accumulated maximum.
Surprisingly, the analysis of RNA samples of a synchronised Med4 culture collected over 24 hours showed a rhythmic accumulation of Yfr7 with different minima and maxima for both transcript variants (Fig. 3.22c): The 332 nt RNA decreases at the day-night-transition and increases during the night obtaining highest levels in the morning; the shorter 220 nt RNA accumulates to a maximum at noon and increases to a minimum around midnight.

The differential expression of both Yfr7 transcripts during growth within ten days indicates a well regulated transcriptional response to likely limiting conditions by peaking during entry into stationary phase at day six (long transcript) or within stationary phase at day eight (short transcript), similar to the situation in *Bacillus subtilis*.
Intriguingly, an effect of the daytime on Yfr7 accumulation was also identified: Again both

Yfr7 variants were peaking differentially within a rhythmic expression pattern. Thus, the cyanobacterial ortholog of 6S RNA, Yfr7, could be involved in the regulation of cellular processes depending on light/darkness and therefore on the circadian rhythm.

Knock-out constructs and *Synechococcus* WH 8102 mutants

Taking the new findings about Yfr7 and 6Sa into account, it appeared very tempting to get insights into the precise function of the cyanobacterial 6S ortholog and to demonstrate that it is also functionally related to the γ-proteobacterial 6S RNA. Thus, the *yfr7* gene of WH 8102 was targeted for a single-cross-over insertion. The transformation was successful as the obtained WH 8102 transconjugants were growing on the chosen antibiotic, kanamycin. Thus, the transformants were plated to get single colonies of one mutant variant.

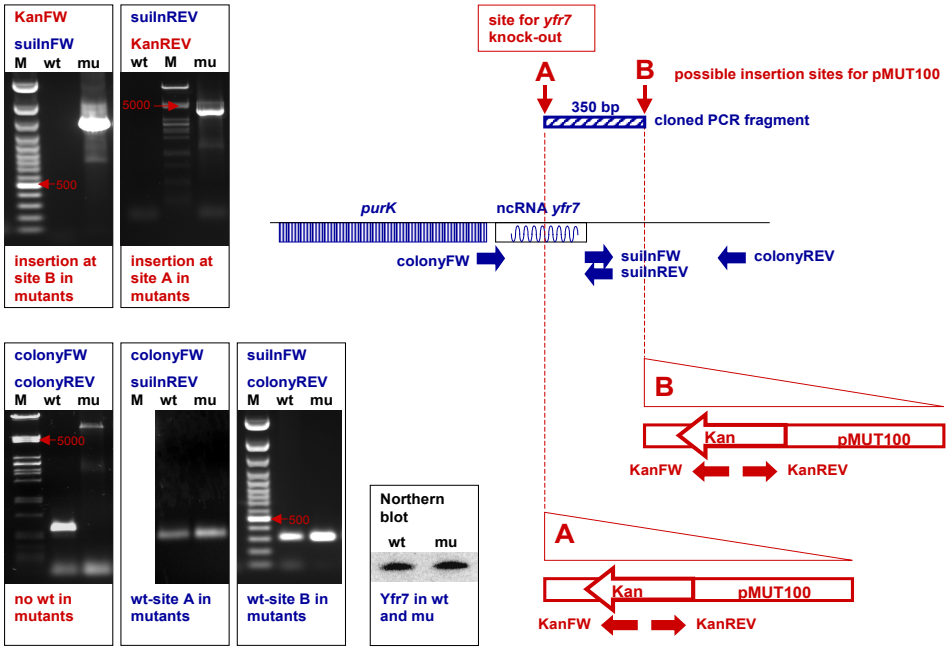


Figure 3.23: Knock-out construct for Yfr7 in *Synechococcus* WH 8102 and analysis of mutants. The *yfr7* gene of WH 8102 was targeted for a single-cross-over insertion: A 350 bp DNA fragment was cloned into pMUT100 knock-out vector; mutants were generated by bi-parental mating and insertion of the knock-out construct into the WH 8102 genome at site A (*yfr7* destruction) or site B (no *yfr7* destruction). Mutant (mu) and wild-type (wt) cultures were analysed by PCR and Northern blot hybridisation, which revealed a mixed DNA population without a successful sorting out of a *yfr7* knock-out.

Further on, the insertion was clearly verified by PCR, but it was not possible to isolate Yfr7-deficient cells up to now: In any case, although an insertion of the foreign construct (including the kanamycin-resistance cassette) into the targeted genomic location was demonstrated (Fig. 3.23), it was not possible to isolate a pure knock-out strain containing only one type of insertion. Additionally, hybridisation to Yfr7 on Northern

blots loaded with total RNA of wild type and mutants revealed signals virtually identical in abundance and molecular weight (Fig. 3.23). Collectively, these data suggest that a mixed mutant population is still existing or that a single WH 8102 cell harbours more than one copy of DNA, resulting in the expression of sufficient amounts of Yfr7. Similar results were observed for knock-out experiments for Yfr1 in WH 8102 (data not shown). Thus, the gene disruption technique based on a single cross-over event might not be the most appropriate one for short target genes. Since a minimum of 300 bp is needed for a successful crossing over, the disrupting plasmid can integrate within or outside the target gene. Therefore, screening for successful gene disruption (by PCR) is very time consuming. However, since now Yfr1 and Yfr7/6Sa are identified in freshwater cyanobacteria like *Synechocystis* PCC 6803 for which molecular tools are more developed than for marine cyanobacteria, the function of these highly interesting RNAs can now be studied in freshwater strains.

3.5.6 Excursus to *Synechocystis* PCC 6803: A *cis*-encoded antisense RNA for *isiA*

Studying the gene expression of *isiA*, encoding the chlorophyll binding protein IsiA, in *Synechocystis* by the group of Dr. Annegret Wilde, a small RNA was detected that is transcribed from this region. More detailed RNA analyses in the context of this work showed that the accumulation of this RNA is inversely regulated to the mRNA level of the stress-inducible gene *isiA* (Fig. 3.24). Therefore, this newly identified RNA, was named IsrR for Iron Stress Repressed RNA. Strand-specific oligonucleotide probes as well as RACE experiments clearly identified this transcript as an antisense RNA transcribed from the complementary strand of *isiA* and from an own promoter (Fig. 3.25a,b).

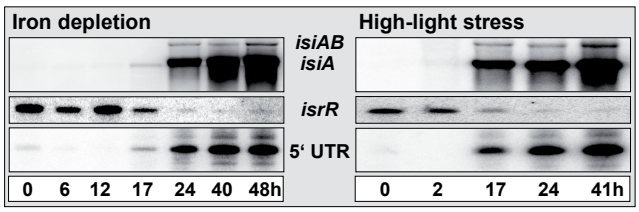


Figure 3.24: Accumulation of different transcripts from the *isiAB* region under iron limitation (left panel). Cells were cultivated under iron replete conditions to the logarithmic phase of growth (0 h), washed in iron-free media and grown further in iron deplete conditions for 6 to 48 h. The *isiAB* dicistronic precursor transcript, the *isiA* mRNA and a transcript in the 5‘ UTR region of *isiA* are induced under iron deplete conditions. The antisense RNA IsrR is inversely accumulated: IsrR is present under iron replete conditions and disappears with increasing time of iron limitation. Induction of the mRNA for *isiA* and *isiAB* and of the 5’ UTR transcript during high-light conditions (right panel). Cells were cultivated under normal light conditions (0 h) and then transferred to high light for 2 to 41 h. The *isiAB*, *isiA* and 5‘ UTR are induced with increasing time of incubation under high light whereas the amount of IsrR decreases. The short transcripts were detected on Northern blots prepared from 10 % polyacrylamide gels using strand-specific oligonucleotides for IsrR and a 5‘ UTR DNA fragment as radiolabelled probes. The *isiA* and *isiAB* transcripts were separated on 1.3 % formaldehyde-agarose gels and hybridised with an *isiA*-specific DNA fragment.

The length of IsrR and by that the precise nucleotide sequence was determined by RACE experiments targeting the 5’ and 3’ ends. The position in the genome extends from 1518034 to 1518210 on the complementary strand. Thus, the 177 nt long IsrR is located in the middle of the annotated coding region of *isiA* (Fig. 3.25a,d). The determined IsrR RNA sequence can be predicted to fold into two extended stem regions each finishing with a terminal loop (Fig. 3.25c).
In addition, another short transcript of about 160 nt is originating from the *isiAB* region, representing a major part of the 5’ untranslated leader region (5’ UTR) of the *isiA* gene. This 5’ UTR transcript was observed before (Vinnemeier et al., 1998), but without assigning specific function to it.
The newly identified antisense RNA IsrR is present in high amounts under iron replete conditions but absent under iron limitation. In contrast, *isiA* mRNA was not detectable under iron replete conditions but with an increasing concentration after 17 h in iron deplete media (3.24). As the *isiA* gene is known to be responsive to high-light-induced stress as well (Havaux et al., 2005), the response of cells transfered from normal to high light was

tested for 2 to 41 hours. Induced by high light, the amount of *isiAB* dicistronic precursor transcript, *isiA* mRNA and 5‘ UTR transcript are gradually increasing, whereas the IsrR RNA signal is fading away rapidly (Fig. 3.24).

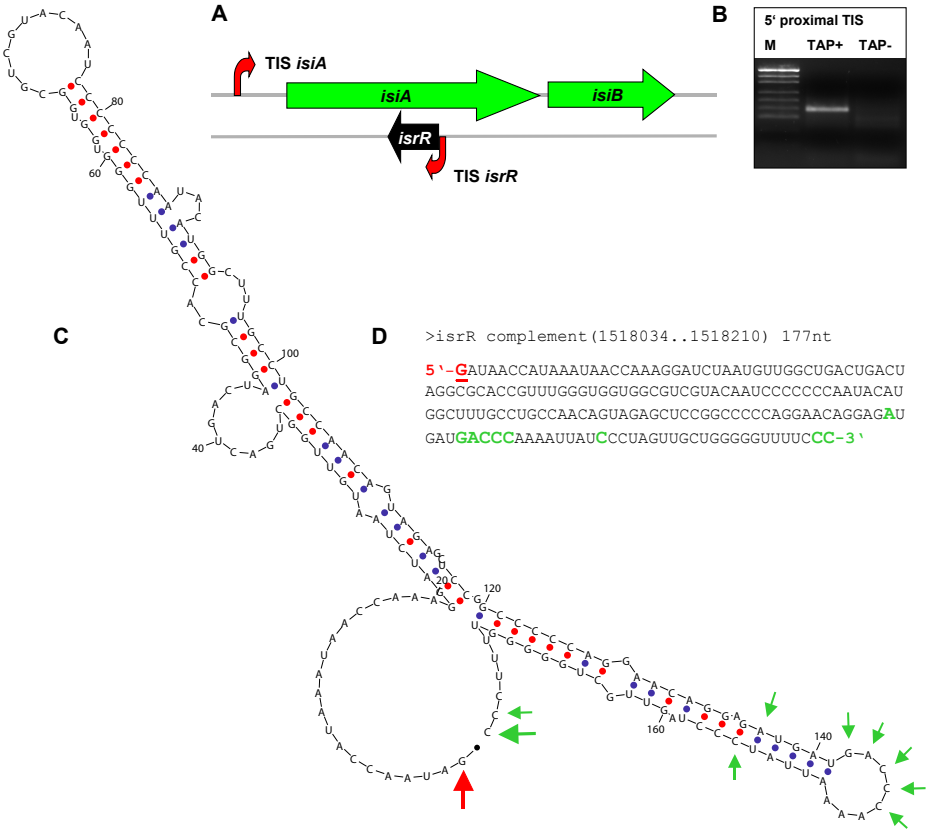


Figure 3.25: Characterisation of an antisense RNA in *Synechocystis* PCC 6803. (A) Location of the *isiA* gene within the *Synechocystis* genome. The long distance between transcription start (red arrow) and translation start site (green arrow) indicates the location of a previously described 5’ UTR transcript in sense orientation (Vinnemeier et al., 1998)). Directly downstream of *isiA* the *isiB* gene is located encoding flavodoxin. (B) The unprocessed 5’ end (lane TAP+) of IsrR detected by 5’ RACE indicating the TIS for *isrR*. (C) RNA secondary structure prediction for IsrR by MFOLD. The red arrow points to the 5’ end of the RNA detected in RACE experiments, the green arrows label the found 3’ ends, whereby the larger green arrow represents the most frequently observed 3’ end. (D) Verified sequence of IsrR: The red letter indicates the 5’ end (TIS), green letters are different 3’ ends.

Thus, at least four different transcripts originate from the *isiAB* genomic locus: The unprocessed *isiAB* precursor transcript, the *isiA* monocistronic transcript, IsrR with maximum accumulation under iron replete and low-light conditions, when *isiA* mRNA is missing, and the 5’ UTR transcript which appears to be co-accumulated with *isiA* mRNA. For each newly identified non-coding RNA it often appears to be challenging on which targets it may act. In this example, the reciprocal relation between the amounts IsrR and *isiA*

mRNA suggests that they could react to each other via a titration mechanism.

Chapter 4

Discussion

4.1 Promoter architecture and transcriptional regulation

A unique comparative dataset is given by the highly related but even so diverse adapted marine cyanobacteria, *Prochlorococcus* Med4, SS120, MIT 9313 and *Synechococcus* WH 8102. The fact, that these four genomes were already sequenced at the beginning of this thesis, offered the great opportunity to study their transcriptional regulation based on sequence comparisons. Firstly, combined experimental and computational promoter studies revealed similar features as known from *E. coli*: The marine cyanobacterial promoter consists of TIS and -10 element, and in the minority of cases also the -35 region is conserved (Vogel et al., 2003a). Additionally, a -60 element occurs for several analysed genes, probably representing another promoter element, which could also be linked to the frequent head to head orientation of genes resulting in an overlapping and often co-regulated (bidirectional) promoter region (Korbel et al., 2004). On the other hand, the sequence composition of spacer regions between conserved promoter elements appears to be more or less random as well as not every comparable genomic region between the four strains possesses conservations in their non-coding parts. Thus, the set of four marine cyanobacterial genomes might offer a reliable genomic distance highlighting conserved sequence elements as biological relevant motifs among orthologous promoter regions within a context of otherwise non-conserved sequence.

Beside putative transcription factor (TF) binding sites for the CRP-like regulator NtcA (Palinska et al., 2000), known to mediate nitrogen control in cyanobacteria, not much was known about the regulatory network for marine cyanobacteria, and genome-wide studies did not exist about co-regulated genes (regulons) controlled by *trans*-acting TFs and their *cis* encoded DNA binding sites. Thereby, a search for additional promoter elements was tempting to get an overview of the transcriptional regulatory potential of these very specialised organisms. Nearly half of all genes in the genomes of marine cyanobacteria are of unknown function. Photosynthesis and an autotroph lifestyle can be assumed to require a set of regulatory circuits that is significantly different from what is known from model organisms such as the intensively studied enterobacteria.

4.1.1 The conserved regulatory potential of four marine cyanobacteria

A phylogenetic footprinting approach for the four sequenced marine genomes was implemented to reveal functionally relevant similarities between promoter regions. Thus, new information was obtained about the fundamentals of transcriptional regulation for the marine cyanobacterial strains each adapted to its particular habitat.

In a first step, a set of orthologous coding regions was calculated resulting in 1293 clusters, which represents a number similar to other BLASTp comparisons of the four genomes, *Prochlorococcus* Med4, SS120, MIT 9313 and *Synechococcus* WH 8102 (Hess, 2004; Dufresne et al., 2005). Keeping in mind, that the total number of coding regions in the smallest of the four genomes, Med4, is 1716, about three-quarter of all Med4 genes belong to this conserved core, whereas only half of the *Synechococcus* WH 8102 genes are represented by these orthologous clusters. Within this set of clusters, only five sigma and 35 putative regulatory factors were found including about 18 potential transcriptional factors, which constitute the predicted core set of regulatory proteins conserved between these four marine cyanobacteria.
This set of regulatory proteins as identified here is mainly limited by the Med4 genome, which is the most compact of any free-living photoautotroph, suggesting that it may possess only that part of the regulatory network of a cyanobacterial cell which is most fundamental and widespread. Therefore, the presented data were expected to be relevant not only for comparison to more closely related marine strains but also to many other cyanobacteria.

About 20 motifs were detected with more than one entry per genome (Appendix, Fig. 2), including sites similar to previously described consensus sequences of the regulators ArsR, LexA and NtcA, found in freshwater cyanobacteria. Focussing on palindromic motifs and following further genome-wide searches of additional entries for the putative binding sites and assignment of their possible regulons, five best motifs were suggested:

(1) An ArsR-like consensus sequence is located within the spacer region of *arsR* and *gap*. However, the *arsBHC* operon that is involved in arsenic sensing and resistance in *Synechocystis* PCC 6803 (Lopez-Maury et al., 2003) was not found in the four marine strains. Thus, the ArsR-like factor here may participate in the regulation of other genes and operons. Indeed, a system identical to *Synechocystis* becomes more unlikely, as the predicted ArsR-like sites appeared upstream of genes like *pstS*, *phoB* (two-component response regulator, phosphate) and *phoR*, thought to be regulated by the amount of phosphate in the cell. As both regulators, PhoB and ArsR, are encoded in each of the four marine genomes studied here, a crosstalk between their regulons might be assumed with the exception of SS120, where *phoB* is missing. On the other hand, a predicted Pho box by Su et al. (2003) was not identified in upstream regions together with this marine ArsR-like motif.
Arsenate is chemically analog to phosphate, and for *Anabaena variabilis* an inhibition of the phosphate transport by intracellular arsenate was already suggested (Thiel, 1988), although the underlying mechanism is still poorly understood. Thus, the repression of certain P-dependent genes in the presence of arsenate mediated by ArsR binding in their promoter regions might represent a possibility.
Interestingly, a minimal cellular P content of *Prochlorococcus* and *Synechococcus* cells

under P-starved conditions is one of their most striking features, indicating very low phosphate requirements (Bertilsson et al., 2003). Even in a non-limiting environment, *Prochlorococcus* was found to retain an intracellular P limitation (Bertilsson et al., 2003). Therefore, one could image, that the regulation of the phosphate uptake and content of these cells has to be very specific. Future experimental efforts may gain further insights into this highly complex regulatory network.

(2) Another palindromic sequence was observed within the upstream region of *ham1* and *csoS1-rbcLS* operon. Additional entries occurred for *petF* and *psbA*, encoding proteins involved in photosynthesis. Thus, this predicted *cis* element might be involved in the regulation of carbon fixation and photosynthetic light reaction. The mRNA level of *psbA* (encoding the reaction centre II subunit D1) was shown to be tightly correlated with light irradiance, with a minimum at night and a maximum at noon (Garczarek et al., 2001) and also *rbcL* expression of the indigenous phytoplankton was greatest in the day and least at night (Pichard and Paul, 1991). A more detailed study, using two cyclostats of *Prochlorococcus* culture, revealed a transcript accumulation of *rbcL* increasing from the beginning of the dark period to a maximum in the early morning and afterwards decreasing sharply down to nearly zero at the end of the light period (Bruyant et al., 2005). Thus, at least the maximum transcript accumulation of *rbcL* and *psbA* appeared to be shifted about several hours, which does not necessarily exclude a co-regulation.
Moreover, a CO_2 response element and the corresponding *trans*-acting factor of the promoter for *rbcLS* in *Synechococcus* sp. PCC7002 were described earlier, which possibly regulate the *rbcLS* transcription in response to CO_2 levels (Onizuka et al., 2002) - a mechanism which is quite similar to the regulation of the *rbcS* promoter in higher plants and which might be assumed for the marine strains, studied here, as well. The RuBisCO operon transcriptional regulator, which is found in all cyanobacteria so far (similar to HTH-, LysR-type Ycf30 encoded on plastids of eukaryotic algae), might mediate this CO_2 response or an additional regulatory circuit.

(3) The putative LexA site found for marine strains is highly similar to the previously described consensus sequences of gram-positive and freshwater cyanobacteria (Mazon et al., 2004b). Furthermore, the LexA regulon predicted here contains several genes known to be active in the SOS response system of bacteria such as *umuC* and *umuD* and especially *recA* and *lexA*, encoding the positive and negative regulator respectively, which might indicate a mechanism surprisingly similar to the SOS system best known from *E. coli* (Walker, 1984).
RACE experiments showed that the predicted binding sites of LexA are located exactly at the transcription initiation site of *lexA*, *umuD* and PMM1427 in Med4, indicating that LexA itself might be negatively autoregulated and could act as the repressor for several other genes.
Although today, there are different functions for LexA discussed in literature (Mazon et al., 2004b; Domain et al., 2004; Gutekunst et al., 2005) and studies about *Synechocystis* (Domain et al., 2004; Gutekunst et al., 2005) raised the question if all cyanobacteria possess an *E. coli*-type SOS regulon, the data obtained during this study for marine cyanobacteria give evidence for a DNA repair system similar to the *E. coli* model.

(4) NtcA is a major regulator for nitrogen control in cyanobacterial cells (Herrero et al., 2001). Those parts of the genome, which are repressed or activated by its presence, constitute the N-regulon. Here, only a small but high-scoring subset of this putative regulon was defined, including genes for major enzymes of nitrogen-metabolising pathways such as *spt,agt* (aminotransferase) and *glnA* (glutamine synthetase) as well as important nitrogen dependent transport systems like *urtABCDE* (urea transporter). The consensus sequence, identified here, harbours additional features besides the often used TGT-N_{10}-ACA motif: The flanking A/T-rich sequences and a conserved TG (or CA) dimer (Vazquez-Bermudez et al., 2002). This more complex motif for marine cyanobacteria corresponds only partly to the profile GTA-N_8-TAC suggested recently (Su et al., 2005). Thereby, only a subset of genes was predicted here compared to the larger NtcA regulons described by Su et al. (2005), in which genes involved in photosynthesis were additionally identified. Nitrogen is an important bio-element and a more global regulation mediated by NtcA might be assumed, that would also affect photosynthesis genes and others. On the other hand, it remains to be proven experimentally, if NtcA is directly involved in a general response of such a large set of genes or if the global influence observed during nitrogen-starvation includes secondary effects as well.

(5) The last of the five top motifs consists of a 16 bp palindrome. The prediction of its likely regulon revealed several genes with an unknown function indicating a completely new binding site for a new regulon in marine cyanobacteria.

A core set of regulons for marine unicellular cyanobacteria is suggested here for the first time. The comparison of four sequenced genomes gives new insights into the minimum network of transcriptional regulation for strains within the marine ecosystem, but it does also allow drawing conclusions for cyanobacteria in general: Two known regulators, NtcA and LexA, appeared to be conserved over a wider evolutionary distance from freshwater to the group of marine cyanobacteria - from the most primitive unicellular to the filamentously growing complex species. The identification of NtcA and LexA in marine cyanobacteria illustrates how the data set might be utilised for an identification of promoters and regulatory sequences in other cyanobacterial species.
In contrast, other factors like the one recognising the ArsR-like binding site, might have evolved differentially and probably possess new functions and regulons adapted to the marine environment. Further experiments and comparisons with high throughput gene expression data will improve this initial regulatory network.
Moreover, one has to remark that the computational predictions of DNA binding sites made here together with the experimentally tested examples can not serve as the entire proof of their biological function. For this purpose, additional binding studies, e.g. DNA affinity precipitation, DNase I protection or mobility shift assays, as well as detailed mutational analyses of the appropriate promoter regions might follow in the next future.

Of course, this small set of transcription factors can not represent the complete one for each strain, because it is missing several factors and their binding sites, which are not conserved between all four strains. Additionally, the deeper focus on palindromic and highly conserved motifs excluded all regulators, which recognise patterns of different geometries and lower base-pair conservations like Fur (ferric uptake regulator protein) binding sites

and others.
Thus, combinations of varying genome-sets as input for the phylogenetic footprinting analysis might help to generate more focussed information on the specific regulation network of a single strain or a set of few strains adapted to a particular niche in the marine ecosystem - a promising possibility for the very next future as there become more and more marine cyanobacterial genomes sequenced.

4.1.2 Circadian rhythm: clock proteins and other periodosome components

Cyanobacteria are the only prokaryotes which possess circadian oscillations. These 24-hour rhythms are generated basically by three proteins in the cyanobacterial cell: KaiA, KaiB and KaiC (Nakajima et al., 2005), culminating in the formation of a high-molecular-weight complex during the night - the periodosome (Bell-Pedersen et al., 2005). Among the known components of a cyanobacterial clock, two genes appear to be deleted from marine strains. The circadian input kinase, CikA, is missing in all strains, so that only one input pathway might remain, mediated by a recently identified protein, LpdA, sensing the redox state of the cell (Ivleva et al., 2005).
Additionally, the gene encoding the core clock protein KaiA can not be annotated for *Prochlorococcus* strains sequenced so far, though it is an indispensable part of the *kai* operon in related marine *Synechococcus* strains. Only for some genomes of low-light-adapted *Prochlorococcus* strains, a truncated *kaiA* coding region can be detected within the *kaiB-rpl21* spacer region, indicating an evolutionary intermediate between the three- and the two-gene form of the *kai* operon. That is in itself intriguing, because it directly contradicts speculations on the *Prochlorococcus kai* gene operon (and clock) as the primitive evolutionary ancestor to the more complex situation found in *Synechococcus elongatus* (Dvornyk et al., 2003).
It has been shown recently, that the genomes of *Prochlorococcus* cells underwent an intensive evolutionary reduction process (Dufresne et al., 2005). Here, the genome reduction process can even be demonstrated by this example: the step by step deletion of the *kaiA* gene. Thus, a reduced clock and a simplified circadian mechanism might be unsurprising for an even so minimised organism like *Prochlorococcus*. The question remains if in *Prochlorococcus* cells an autonomous pacemaker can exist, ticking without the basic KaiA protein component? For the model organism *Synechococcus* PCC 7942 all three *kai* genes are essential for circadian rhythmicity, and an inactivation of any of them abolishes it (Ishiura et al., 1998). On the other hand, studying mutations of sigma factors revealed that PkaiA and PkaiB are on different regulatory circuits and that the *kaiA* expression can be dramatically altered without changing the fundamental timing mechanism (Nair et al., 2002).
Based on the data presented in this thesis, the transcriptional regulation of the *kai* genes in marine strains differs from the model: A *kaiBC* promoter could not be found for marine *Synechococcus* WH 8102. Furthermore, in *Prochlorococcus* the expression of *kaiB* seems to be coupled to that of the ribosomal gene *rpl21*. One might assume that the reduction of the clock components and the different timing of *kai* gene expression as compared to *Synechococcus* PCC 7942 might result in a simplified circadian clock for marine cyanobacteria.

Further studies of the marine circadian system including its essential protein components are needed to understand all the underlying regulatory mechanisms.

4.1.3 Transcriptional signals of cyanophage P-SSP7 infecting *Prochlorococcus* Med4

The cyanophage P-SSP7 possesses several T7-like genes including the essential RNA polymerase (Sullivan et al., 2005). Additionally, a gene expression order from left to right (Lindell, personal communication) implicates a transcriptional strategy as described for the T7 group of phages. Intriguingly, in a computational study not a single promoter site similar to the well-known and -conserved phage promoters was detected (Chen and Schneider, 2005), which might indicate differences of the transcriptional mechanism for the P-SSP7 genome. An analysis of phage mRNAs on the basis of new experimental data was thought to discover sites of transcription initiation and therefore recognise promoters to solve the question for the unknown transcription signals of RNA polymerase in P-SSP7.

Predictions of promoter and termination sites as well as the experimental analysis of the mRNA 5' ends revealed transcriptional features of P-SSP7, which are partly similar but also different to the T7 model. Transcription of P-SSP7 from left to right demands transcription of the first phage genes (including RNAP) from bacterial promoter sites located near the left end of P-SSP7 DNA. A Med4-like promoter region was predicted upstream of gene 1, although it could not be verified experimentally. The initial transcription by the host polymerase might be stopped at the predicted termination sites between gp12 and gene 3 or upstream of gp1, the gene for the RNA polymerase, which is not in agreement with T7, where the expression of the phage RNA polymerase is essential for the following steps of infection.
On the other hand, a bacterial-like promoter was found upstream of the gene for the phage RNAP, which might indicate the possibility that this important phage gene is transcribed independently of the other P-SSP7 genes by the host RNA polymerase. Consequently, the phage RNAP protein would possibly be present for the transcription of P-SSP7 DNA very early during Med4 infection. Nevertheless, the time order of gene transcription in P-SSP7 remains unclear.
Another similarity to a T7-like transcription was detected by the prediction of a termination site between the capsid protein (gp10) and the tail tubular protein gene (gp11). As a promoter for the latter gene is absent, gp11 has to be transcribed by read-through of the terminator, which is part of the transcriptional strategy in T7 and ensures production of large amounts of the major capsid protein, gp10 (Dunn and Studier, 1983).

RNA samples of P-SSP7 infected *Prochlorococcus* Med4 cells were used in RACE experiments in order to analyse the transcriptional units expressed during the infection cycle. Various 5' ends were identified for cyanophage mRNAs. Among them, one group with signals of maximum strength and stability for the genes 1, 3 and 29 (gp10, capsid protein) revealed a motif conserved between these three sites, which was shown to possess more similarity to an RNase cleavage site than to a promoter site for the phage polymerase. Additionally, inverted repeats are located next to this motif, which might serve as recog-

nition sites for RNase III - a processing mechanism of phage mRNAs described for the T7 infection as well (Dunn and Studier, 1983).
The fact, that the motif was also conserved in two environmental sequence samples (Venter et al., 2004) upstream of orthologous gp10 genes, supports the idea of an essential element for the P-SSP7 phage transcription.
Additional entries of the motif appeared within the host genome next to putative 3' cleavage sites of tRNA and tmRNA. Thus, the identified motif might possess a biological function for both, *Prochlorococcus* Med4 and P-SSP7, in representing a potential cleavage site for their transcription products mediated by an RNase.

The phage encoded genes, *hli* and *psbA*, did not reveal a detectable promoter signal, supporting the suggestion of co-transcription with the essential phage capsid genes (Lindell et al., 2005).
Two host genes of *Prochlorococcus* Med4, PMM0684 and PMM0819, induced during infection (Lindell, unpublished) possess perfect bacterial promoters, verified by RACE signals, which might indicate a response of the host to its infecting phage.

It still remains questionable if the well studied transcriptional mechanism for coliphage T7 could be assigned to the marine phage P-SSP7 one to one. The observation from T7, that promoter and cleavage sites often occur next to each other, might had complicated the experimental detection of *in vivo* transcription initiation sites. Thus, further experiments like an *in vitro* transcription system for P-SSP7 RNA polymerase would be helpful to analyse its promoter recognition sites. Additionally, future studies of RNA processing mechanisms in the host, *Prochlorococcus* Med4, as well as for its infecting phage, P-SSP7, would provide important information for both members of the marine community.

4.2 New families of small RNAs in cyanobacteria

For several eubacteria, especially for *E. coli* as well as for *Vibrio cholerae*, non-coding RNAs are described as essential regulatory factors mediating rapid responses to environmental changes. The wide variety of underlying regulatory mechanisms are still poorly understood. Compared to the immense and growing number of predicted small RNAs, only a few well investigated functional interactions are described for RNA antisense binding to mRNAs, manipulation of proteins or direct sensing of metabolites (riboswitches).
For free-living marine phototrophs such as the cyanobacterial strains investigated in this study, regulatory circuits involving small RNAs were expected as well. However, except for RNase P RNA, scRNA and tmRNA, three RNAs easily to identify, nearly nothing was explored about ncRNA genes in marine as well as other cyanobacteria. Based on a biochemical protocol, a single ncRNA, 6Sa, was identified previously in unicellular freshwater cyanobacteria, *Synechococcus* PCC 6301 and *Synechocystis* PCC 6803 (Watanabe et al., 1997). In addition, mapping of transcriptional units within the gas vesicle operon of *Calothrix*, a filamentous and heterocysts forming algae, identified a single antisense transcript (Csiszar et al., 1987). Recently, a *cis* antisense RNA was found in the nitrogen-fixing cyanobacterium *Anabaena* sp. PCC 7120 interfering with *furA* transcript translation (Hernandez et al., 2006).

It was already shown for promoter and phylogenetic footprinting analyses that the genomes of *Prochlorococcus* SS120, MIT 9313, Med4 and *Synechococcus* WH 8102 provide a unique dataset for comparative studies. These small, even smallest cyanobacterial genomes differ by several hundred genes from each other, although most of the operons and gene clusters present in more than a single genome are co-linear (Dufresne et al., 2003; Palenik et al., 2003; Rocap et al., 2003). In a first genome-wide and systematical screen for ncRNAs in marine cyanobacteria 17 ncRNAs were detected in Med4, SS120, MIT 9313 and WH8102. With a focus on *Prochlorococcus* Med4 the presence of new non-coding RNAs in the group of marine unicellular cyanobacteria was analysed experimentally.

4.2.1 Computational screening in marine genomes

Genes encoding functional RNAs are notoriously difficult to predict, because of their low conservation in primary sequence. Therefore, in cooperation with Philip Kensche a comparative computational approach based on sequence and structure conservation was implemented for the cyanobacterial genome set (Kensche, 2004), including an algorithm that was recently introduced for the identification of eukaryotic ncRNAs (Washietl and Hofacker, 2004). In view of the daily growing number of microbial genome sequences, comparative approaches will become more and more possible and required as well.
Only highest scoring candidates of these predictions were analysed further, detecting several previously unknown ncRNAs as well as other elements that function at the RNA level. By comparison to literature and experimental data, the predicted list of high-scoring candidates contained a very low rate of true negatives. This indicates that the employed method is very efficient in finding microbial ncRNAs and other RNA elements. Therefore, it can be suggested to exploit this algorithm to further related genome sets predicting ncRNA for cyanobacteria or other microbes as well.

One has to keep in mind, that the 17 ncRNAs detected here in Med4, SS120, MIT 9313 and WH8102 are only a part of the total ncRNA population present in these species. That could be caused by the restriction to intergenic regions. Thus, ncRNAs were missed that reside within annotated regions including the whole class of antisense RNAs, encoded complementary to their target. Additionally, incorrect annotations reduced the number of predicted sequences, like in the case of *yfr7*, which is located in a region complementary to a reading frame in two of the genomes investigated here.

The presence of new non-coding RNAs in the group of marine unicellular cyanobacteria focussed on *Prochlorococcus* Med4 may present only the tip of the iceberg. Several more ncRNA candidate genes were predicted in the two relatively larger genomes of WH 8102 and MIT 9313 but still await experimental testing. An overview of the candidate regions identified by this screen is given in the Appendix, Table 3 and a summary of the experimentally confirmed new ncRNAs is presented in Table 3.5. In addition to small ncRNAs genes, the computational results indicated the presence of conserved secondary structure elements belonging to untranslated upstream regions of several ribosomal protein operons. Thus, autogenous control mechanisms for the expression of ribosomal genes similar to enterobacteria (Lindahl and Zengel, 1986; Zengel and Lindahl, 1994) may exist in marine

cyanobacteria as well. The four genomes analysed here represent only a tiny part of the unmeasured and diverse oceanic gene pool, and the existence of numerous ncRNA genes in newly sequenced genomes as well as in environmental samples (Venter et al., 2004) can be predicted.

4.2.2 Non-coding RNAs of *Prochlorococcus* Med4 and their orthologs

The analysis of four strains of *Prochlorococcus* and *Synechococcus* revealed an interesting set of structural RNA elements. Especially the ncRNAs found in Med4 and SS120 may be of considerable importance, as these strains underwent a strong genome reduction including deletion of the *hfq* gene. The larger genomes of WH 8102 and MIT 9313 contain an *hfq* gene, whose product has been shown to be essential for the activity of small regulatory RNAs in enterobacteria (Valentin-Hansen et al., 2004). It is likely that, together with *hfq*, several ncRNA genes have been deleted during the evolution of the *Prochlorococcus* group towards a minimal genome. Thus, the ncRNAs still remaining in an organism like Med4 must have been subject to strong positive selection and may act independently of Hfq. On the other hand, Hfq binding is not essential for the functionality of every ncRNA, as only 30 % of investigated ncRNAs in *E. coli* have been shown to be bound by Hfq (Zhang et al., 2003).

Evidence of function of the ncRNAs described here may arise from the comparison of expression patterns, structures as well as genomic location, and from the presence or absence of a given ncRNA gene in the different strains:
(1) The Yfr1 RNA might be dispensable for growth at greater depths, because its gene is clearly absent from the ultra low-light-adapted SS120 but present at the identical genome location (upstream *trxA*) in the other three marine strains. Surprisingly, it also appeared to be conserved in freshwater cyanobacteria, *Synechococcus* PCC 7942, *Thermosynechococcus elongatus* and *Synechocystis* PCC 6803, harbouring a highly similar secondary structure with two terminal loops separated by an unpaired region that contains a CA dinucleotide repeat. These highly conserved features might represent essential functional elements of Yfr1.
(2) Yfr2 through Yfr5 have several entries in Med4 and WH 8102, which might be assumed for other strains as well, and are in length and the degree of mutual identity similar to four ncRNAs implicated in quorum sensing in *Vibrio* species (Lenz et al., 2004).
(3) The longer RNA Yfr6 seems to be restricted to the smallest marine genomes, as it was only found in Med4 and SS120. Its very short half-life and a conserved short reading frame represent unique properties compared to the other Med4 ncRNAs analysed experimentally.
(4) The marine Yfr7 RNA was shown to represent an orthologue of 6Sa RNA (*ssaA* gene product) of freshwater and other cyanobacteria as well as for the 6S RNA known from γ-proteobacteria; described in detail in the next section.

Consequently, the ncRNAs identified here may constitute important regulatory or structural components not only of a free-living marine cyanobacterium but also for the whole cyanobacterial radiation. Up to now, the detailed regulatory mechanisms of the detected ncRNAs are unclear, and a genetic manipulation system, especially for such short genes,

is still a challenging task for marine strains. Thus, the ncRNAs which possess orthologues over a wider phylogenetic distance might be functionally analysed directly in these cyanobacteria for which well-established genetic tools exist, as the freshwater strains *Synechococcus* PCC 7942 or *Synechocystis* PCC 6803. Therefore, promising candidates are Yfr1 and Yfr7/6Sa, which are likely present in all cyanobacteria.
Another possibility would be the transfer of the ncRNA as well as the putative target(s) to an appropriate host or model organism to study the mode of ncRNA interaction, often an antisense mechanism. For unicellular marine cyanobacteria, *Synechococcus* WH 7803 might become such a model: There is a genetic manipulation system described (Brahamsha, 1996) and its genome analysis has almost completely been finished (Partensky, 2006). Additionally, whole genome microarrays for WH 7803 are in progress (SynChips project of the European network Marine Genomics, PI: W. Hess), which will allow high-through-put experiments for putative ncRNA-deficient strains in the future.

4.2.3 Ubiquitous presence of 6S RNA in cyanobacteria and clues about its function

In an experimental and detailed computational analysis the marine Yfr7 RNA was identified as the orthologue of 6Sa RNA (*ssaA* gene product), described and partly annotated already for several freshwater and other cyanobacteria. Additionally, a recently published study of Barrick et al. (2005) detected 6Sa computationally in all cyanobacterial strains representing the orthologue of 6S RNA known from γ-proteobacteria. Unfortunately, genomic annotations of 6Sa RNA (*ssaA*) in cyanobacteria are often on the incorrect strand. The transcription of the *ssaA* gene from the complement strand as annotated was demonstrated for *Thermosynechococcus elongatus*, *Anabaena* PCC 7120 and *Gloeobacter violaceus* PCC 7421 in this study; for *Synechococcus* PCC 6301 and *Synechocystis* PCC 6803 the annotated genome location of *ssaA* was shown to be correct.

E. coli 6S RNA binds to the σ^{70} polymerase holoenzyme to globally regulate gene expression in response to the shift from exponential growth to stationary phase. The inhibition of transcription occurs by direct competition for promoter DNA binding to Eσ^{70}, the house-keeping form of RNA polymerase. Thereby, 6S RNA is mimicking the structure of a DNA template in an open promoter complex (Barrick et al., 2005; Trotochaud and Wassarman, 2005). The highly stable secondary structure of 6S RNA essential for this mechanism, containing a single-stranded central bulge within a highly double-stranded molecule (Trotochaud and Wassarman, 2005), was also predicted for the Yfr7/6Sa RNA identified here. Thus, at least a structural ortholog of γ-proteobacterial 6S RNA is conserved within the whole cyanobacterial lineage.
In *E. coli* the 6S RNA is mainly active under stress conditions, like nutrient limitation, and an accumulation can be observed during entry into stationary phase. To get insights into the situation in cyanobacterial strains as well, the growth-phase dependent expression of two transcripts of the Med4 6S RNA homolog have been identified in this study. One transcript of the Med4 6S RNA (220 nt) reaches maximal accumulation at high cell densities. The second, 332 nt long and processed transcript is peaking earlier in Med4, probably during entry into stationary phase. The existence of two differentially expressed

6S RNAs, 6Sa and 6Sb, was recently shown for *Bacillus subtilis*, where 6Sb is the ortholog to *E. coli* 6S, and 6Sa has functionally diverged (Barrick et al., 2005). Although both transcripts of Med4 (and other high-light-adapted strains) originate from only one genomic locus, the different growth-phase dependent expression of Yfr7/6S transcripts in Med4 may also indicate a fine tuning of the transcriptional response to nutrient limitation.

Additionally, a secondary effect of 6S RNA in the activation of σ^S-dependent promoter transcription was observed in *E. coli* cells (Trotochaud and Wassarman, 2004) mediated by the general stress response factor σ^S, which is the only group 2 σ factor in *E. coli*. In contrast, cyanobacteria likely possess a more complex regulatory network of group 2 σ factors. They harbour multiple group 2 σ factors, so that a single orthologous gene can not be assigned to the enterobacterial stationary phase σ factor, σ^S. It was suggested that these multiple group 2 σ factors are differentially active during the day, transmitting the circadian rhythm generated by the cyanobacterial clock (Nair et al., 2002; Ditty et al., 2003). Interestingly, a time-shifted, rhythmic expression pattern exists between both Med4 Yfr7/6S transcripts, which was demonstrated using RNA samples from synchronised cultures.
Therefore, it was of great interest to find out if the cyanobacterial Yfr7/6S RNA is also functionally related to the γ-proteobacterial one. A common experimental approach to achieve more information about the gene of interest is the construction of its knock-out strain. As there was a transformation system already tested for marine *Synechococcus*, the *yfr7/ssaA* gene of WH 8102 was targeted for a single-cross-over insertion. Although the insertion was found to be successful, it was not possible to isolate Yfr7/6S-deficient cells, up to now. But the widespread occurrence of this ncRNA opens additional exciting opportunities to test the function of 6S directly in cyanobacteria like *Synechocystis* PCC 6803, which are easier and faster to manipulate.

4.2.4 A novel antisense RNA to *isiA* mRNA in *Synechocystis* PCC 6803

In cyanobacteria, the iron stress induced gene A, *isiA*, encodes a chlorophyll binding protein homologous to the photosystem II (PSII) protein PsbC (CP43). High attention was paid to IsiA (CP43') when it was detected to form a giant alternative antennae ring around the photosystem I (PSI) trimer under iron limitation (Bibby et al., 2001a; Boekema et al., 2001). However, the precise role and regulation of *isiA* gene expression has remained enigmatic as it is also induced under high light, salt and oxidative stress (Vinnemeier et al., 1998; Jeanjean et al., 2003; Yousef et al., 2003; Havaux et al., 2005). It was suggested that IsiA may serve as a stress-inducible antennae in the dissipation of excess light energy or as storage for excess chlorophyll a (Cadoret et al., 2004; Ihalainen et al., 2005).

Here, the transcription of an antisense RNA, IsrR (Iron stress repressed RNA) from the *isiA* complement strand in the freshwater cyanobacterium *Synechocystis* PCC 6803 was demonstrated, that is inversely regulated to the accumulation of its host gene's mRNA. Thus, at least four different transcripts were shown to originate from the *isiAB* genomic locus: The unprocessed *isiAB* precursor transcript, the *isiA* monocistronic transcript, IsrR with maximum accumulation under iron replete and low-light conditions, when *isiA* mRNA is missing, and the 5' UTR transcript which appears to be co-accumulated with

isiA mRNA.
For each newly identified non-coding RNA it often appears challenging to find out to which targets it may act. In this example, the reciprocal relation between the amounts IsrR and *isiA* mRNA suggests that they could react to each other targeting both for a degradation mechanism. An antisense RNA that is degraded together with its target makes a perfect reversible switch to respond to environmental changes. Indeed, also in the case of *trans*-encoded ncRNAs with imperfect complementarity, a growing number of examples indicate that they can target specific mRNAs for degradation by RNase III. Examples are the IstR RNAs in *E. coli* (Vogel et al., 2004) and the *spa* mRNA in *Staphylococcus aureus* (Huntzinger et al., 2005). Thus, the formation of ncRNA-protein complexes containing RNase III, and in some cases RNase E (Morita et al., 2005), could be a general way by which mRNAs are destabilised by small RNAs in bacteria in a similar way as in higher organisms (Morita et al., 2005).

Depending on the respective stress, different levels of *isiA* induction were observed: A response to high light or iron depletion began after about 17 hours with increasing amounts of *isiA* mRNA until at least 48 hours, whereas a very rapid but short burst of expression occurred as a response to H_2O_2 treatment. This underscores the well decided answer of the *Synechocystis* cell to the chosen conditions and might refer to the underlying mechanism of *isiA* induction: The triggering factor for a highly induced *isiA* expression could be a redox stress over a specified period, which orders the cell to arrange the PSI-IsiA supercomplexes. Interesting future experiments might be the variation of the amount and period for the stress trigger, high light or H_2O_2.

Together, these results suggested a stringent regulatory mechanism acting on the expression of *isiA* at the RNA level, at least. The coupled degradation of IsrR/*isiA* might constitute a perfectly timed reversible switch to respond to environmental changes in iron concentrations, redox conditions and possibly other stresses (Duehring et al., 2006).

The costly expression of giant amounts of *isiA* mRNA under stress conditions and the later high level of IsiA protein arranging into a large alternative antennae ring might indicate a reason why the cell utilises this special antisense mechanism as an additional level of regulation for *isiA* and likely other genes, whose expression is only advantageous for the cell under defined conditions. Moreover, the latest experiments in *Synechocystis* (performed by Ulf Duehring (Duehring et al., 2006)) provided further evidence that the novel IsrR-mediated regulatory mechanism affects not only the *isiA* mRNA abundance but also the assembly of IsiA-PSI supercomplexes as shown in Figure 4.1.

IsrR is the first non-coding RNA regulating a component involved in oxygenic photosynthesis. The fact, that it is located over its full length complementary to a protein-coding region encoding its likely target, might open a new view on bacterial regulatory RNAs, which have been thought so far to be transcribed mainly from intergenic regions. Thus, a future focus to protein-coding regions, especially for genes underlying stringent but unknown regulatory mechanisms, can be suggested for the bacterial genomes. Further on, the designed opening of a well-defined expression window for a mRNA target by a *cis*-acting RNA is more alike mechanisms reported for eukaryotes. Again, a regulatory feature might

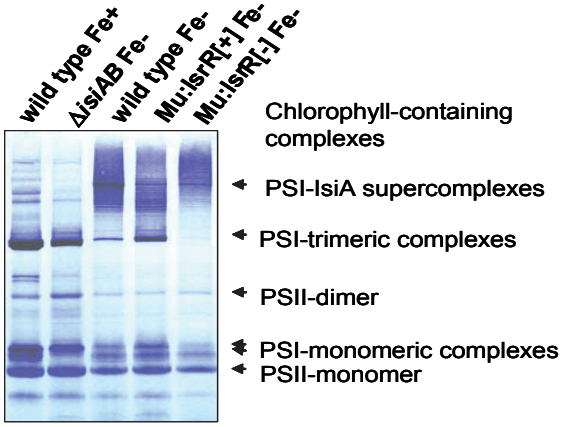


Figure 4.1: Influence of IsrR on the accumulation of trimeric PSI and PSI-IsiA supercomplexes (Duehring et al., 2006). Membrane protein complexes isolated from thylakoids of *Synechocystis* PCC 6803 wild type, an *isiAB* deletion mutant, the IsrR overexpressor (Mu:IsrR[+]) mutant, and cells with reduced amounts of IsrR (Mu:IsrR[-]). Thylakoid membranes were isolated before (Fe+) and after 2 days of iron starvation (Fe-) and separated by Blue-Native-PAGE. Coomassie-stained sections of the Blue-Native-gel are shown. Chlorophyll containing protein complexes are indicated on the right. PS, photosystem.

not be invented the first time for a higher organism and is likely existing for prokaryotes as well.

4.3 Conclusion and outlook

Computational prediction algorithms are often limited in detecting biological relevant signals (Hu et al., 2005). Thus, a trend of recent motif discovery algorithms is to incorporate additional information to improve the prediction accuracy. For this purpose, comparative studies are often implemented to identify conservations between genomes, which are assumed to represent essential elements of the group of organisms investigated. In this study about marine cyanobacterial genomes, it was successfully demonstrated that the comparison of four related genomes revealed so far unknown motifs for transcription factors and even new genes for non-coding RNAs, which might have never been detected within a single genome.

The global importance of the marine ecosystem including the major group of oxygen-producing *Synechococcus*/*Prochlorococcus* becomes indisputable these days. For these organisms, a giant gene pool exists in the ocean, which is mixed up by viruses and possibly other mechanisms of lateral gene transfer still to be discovered. Only a tiny fraction of the marine biodiversity is being cultivated in the laboratory, and from an even smaller number genome sequences are available for studies like this.

One has to keep in mind the fact, that one genome sequence resembles only a snapshot for a single organism. In reality, this organism is part of a larger community which is constantly changing to adapt to variations in its environment. Thus, a suggestive approach might be analysing many related bacteria in parallel - the microbial "pan-genome" consisting of

a core genome shared by all members plus a dispensable genome consisting of partially shared and strain-specific genes (Tettelin et al., 2005) to better understand evolutionary processes. Thereby, RNA genes as well as transcription factors constitute very informative sequences to connect the genome information to the conditions set by the respective environmental constraints. Additionally, phylogenetic analysis of both, DNA elements and the proteins that recognise them, might provide new conclusions about their evolutionary history as demonstrated for LexA (Mazon et al., 2004a) and T7 RNA polymerase (Chen and Schneider, 2005).

Environmental sequence data are now available (Venter et al., 2004) and ongoing genome projects are providing a fast-growing number of sequences, although unique genes will continue to be identified even after sequencing hundreds of genomes (Tettelin et al., 2005). Thus, new ways need to be explored to gain deeper insights into complex networks such as the community of marine cyanobacteria and their phages.

Bibliography

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–10.

Altuvia, S., Weinstein-Fischer, D., Zhang, A., Postow, L., and Storz, G. (1997). A small, stable RNA induced by oxidative stress: role as a pleiotropic regulator and antimutator. *Cell*, 90(1):43–53.

Altuvia, S., Zhang, A., Argaman, L., Tiwari, A., and Storz, G. (1998). The Escherichia coli OxyS regulatory RNA represses fhlA translation by blocking ribosome binding. *EMBO J*, 17(20):6069–75.

Argaman, L. and Altuvia, S. (2000). fhlA repression by OxyS RNA: kissing complex formation at two sites results in a stable antisense-target RNA complex. *J Mol Biol*, 300(5):1101–12.

Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E., Margalit, H., and Altuvia, S. (2001). Novel small RNA-encoding genes in the intergenic regions of Escherichia coli. *Curr Biol*, 11(12):941–50.

Asato, Y. (2003). Toward an understanding of cell growth and the cell division cycle of unicellular photoautotrophic cyanobacteria. *Cell Mol Life Sci*, 60(4):663–87.

Asayama, M., Imamura, S., Yoshihara, S., Miyazaki, A., Yoshida, N., Sazuka, T., Kaneko, T., Ohara, O., Tabata, S., Osanai, T., Tanaka, K., Takahashi, H., and Shirai, M. (2004). SigC, the group 2 sigma factor of RNA polymerase, contributes to the late-stage gene expression and nitrogen promoter recognition in the cyanobacterium Synechocystis sp. strain PCC 6803. *Biosci Biotechnol Biochem*, 68(3):477–87.

Axmann, I., Kensche, P., Vogel, J., Kohl, S., Herzel, H., and Hess, W. (2005). Identification of cyanobacterial non-coding RNAs by comparative genome analysis. *Genome Biol*, 6(9):R73.

Axmark, D., Larsson, A., and Widenius, M. (2006). MySQL AB: The world's most popular open source database. http://www.mysql.com/.

Barrick, J., Sudarsan, N., Weinberg, Z., Ruzzo, W., and Breaker, R. (2005). 6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter. *RNA*, 11(5):774–84.

Barry, G., Squires, C., and Squires, C. (1980). Attenuation and processing of RNA from the rplJL–rpoBC transcription unit of Escherichia coli. *Proc Natl Acad Sci U S A*, 77(6):3331–5.

Bell-Pedersen, D., Cassone, V., Earnest, D., Golden, S., Hardin, P., Thomas, T., and Zoran, M. (2005). Circadian rhythms from multiple oscillators: lessons from diverse organisms. *Nat Rev Genet*, 6(7):544–56.

Bensing, B., Meyer, B., and Dunny, G. (1996). Sensitive detection of bacterial transcription initiation sites and differentiation from RNA processing sites in the pheromone-induced plasmid transfer system of Enterococcus faecalis. *Proc Natl Acad Sci U S A*, 93(15):7794–9.

Berriman, M. and Rutherford, K. (2003). Viewing and annotating sequence data with Artemis. *Brief Bioinform*, 4(2):124–32.

Bertilsson, S., Berglund, O., Karl, D., and Chisholm, S. (2003). Elemental composition of marine prochlorococcus and synechococcus: Implications for the ecological stoichiometry of the sea. *Limnology and Oceanography*, 48:1721–31.

Bibby, T., Nield, J., and Barber, J. (2001a). Iron deficiency induces the formation of an antenna ring around trimeric photosystem I in cyanobacteria. *Nature*, 412(6848):743–5.

Bibby, T., Nield, J., Partensky, F., and Barber, J. (2001b). Oxyphotobacteria. Antenna ring around photosystem I. *Nature*, 413(6856):590.

Birge, E. (2000). *Bacterial and Bacteriophage Genetics (4th edn)*. Springer Verlag, New York.

Birkey, S., Liu, W., Zhang, X., Duggan, M., and Hulett, F. (1998). Pho signal transduction network reveals direct transcriptional regulation of one two-component system by another two-component regulator: Bacillus subtilis PhoP directly regulates production of ResD. *Mol Microbiol*, 30(5):943–53.

Blanco, A., Sola, M., Gomis-Ruth, F., and Coll, M. (2002). Tandem DNA recognition by PhoB, a two-component signal transduction transcriptional activator. *Structure*, 10(5):701–13.

Blattner, F., Plunkett, 3rd, G., Bloch, C., Perna, N., Burland, V., Riley, M., Collado-Vides, J., Glasner, J., Rode, C., Mayhew, G., Gregor, J., Davis, N., Kirkpatrick, H., Goeden, M., Rose, D., Mau, B., and Shao, Y. (1997). The complete genome sequence of Escherichia coli K-12. *Science*, 277(5331):1453–74.

Boekema, E., Hifney, A., Yakushevska, A., Piotrowski, M., Keegstra, W., Berry, S., Michel, K., Pistorius, E., and Kruip, J. (2001). A giant chlorophyll-protein complex induced by iron deficiency in cyanobacteria. *Nature*, 412(6848):745–8.

Brahamsha, B. (1996). A genetic manipulation system for oceanic cyanobacteria of the genus Synechococcus. *Appl Environ Microbiol*, 62(5):1747–51.

Bruyant, F., Babin, M., Genty, B., Prasil, O., Behrenfeld, M., Claustre, H., Bricaud, A., Holtzendorff, J., Koblizek, M., Garczareck, L., and Partensky, F. (2005). Diel variations in the photosynthetic parameters of prochlorococcus strain pcc 9511: combined effects of light and cell cycle. *Limnology and Oceanography*, 50:850–63.

Bryant, D. (2003). The beauty in small things revealed. *Proc Natl Acad Sci U S A*, 100(17):9647–9.

Bulyk, M. (2003). Computational prediction of transcription-factor binding site locations. *Genome Biol*, 5(1):201.

Cadoret, J., Demouliere, R., Lavaud, J., van Gorkom, H., Houmard, J., and Etienne, A. (2004). Dissipation of excess energy triggered by blue light in cyanobacteria with CP43' (isiA). *Biochim Biophys Acta*, 1659(1):100–4.

Cao, Q. and Kibbe, W. (2006). Oligonucleotide Properties Calculator. http://www.basic.northwestern.edu/biotools/oligocalc.html.

Chen, S., Lesnik, E., Hall, T., Sampath, R., Griffey, R., Ecker, D., and Blyn, L. (2002). A bioinformatics based approach to discover small RNA genes in the Escherichia coli genome. *Biosystems*, 65(2-3):157–77.

Chen, Z. and Schneider, T. (2005). Information theory based T7-like promoter models: classification of bacteriophages and differential evolution of promoters and their polymerases. *Nucleic Acids Res*, 33(19):6172–87.

Chisholm, P. (2006). CHISHOLM LAB: Protocols, natural seawater-based PRO99 medium. http://web.mit.edu/chisholm/www/protocols/.

Chisholm, S., Frankel, S., Goericke, R., Olson, R., Palenik, B., Waterbury, J. B., West-Johnsrud, L., and Zettler, E. (1992). Prochlorococcus marinus nov. gen. nov. sp.: an oxyphototrophic marine prokaryote containing divinyl chlorophyll a and b. *Archives of Microbiology*, 157:297–300.

Chisholm, S., Olson, R., Zettler, E., Goericke, R., Waterbury, J., and Welschmeyer, N. (1988). A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature*, 334:340–3.

Cliften, P., Hillier, L., Fulton, L., Graves, T., Miner, T., Gish, W., Waterston, R., and Johnston, M. (2001). Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res*, 11(7):1175–86.

Crooks, G., Hon, G., Chandonia, J., and Brenner, S. (2004). WebLogo: a sequence logo generator. *Genome Res*, 14(6):1188–90.

Crooks, G., Hon, G., Chandonia, J., and Brenner, S. (2006). WebLogo. www.bio.cam.ac.uk/cgi-bin/seqlogo/logo.cgi.

Csiszar, K., Houmard, J., Damerval, T., and Tandeau de Marsac, N. (1987). Transcriptional analysis of the cyanobacterial gvpABC operon in differentiated cells: occurrence of an antisense RNA complementary to three overlapping transcripts. *Gene*, 60(1):29–37.

Curtis, S. (1987). Genes encoding the beta and epsilon subunits of the proton-translocating ATPase from Anabaena sp. strain PCC 7120. *J Bacteriol*, 169(1):80–6.

di Bernardo, D., Down, T., and Hubbard, T. (2003). ddbRNA: detection of conserved secondary structures in multiple alignments. *Bioinformatics*, 19(13):1606–11.

Ditty, J., Williams, S., and Golden, S. (2003). A cyanobacterial circadian timing mechanism. *Annu Rev Genet*, 37:513–43.

Domain, F., Houot, L., Chauvat, F., and Cassier-Chauvat, C. (2004). Function and regulation of the cyanobacterial genes lexA, recA and ruvB: LexA is critical to the survival of cells facing inorganic carbon starvation. *Mol Microbiol*, 53(1):65–80.

Douglas, S. (1998). Plastid evolution: origins, diversity, trends. *Curr Opin Genet Dev*, 8(6):655–61.

Dubchak, I. and Frazer, K. (2003). Multi-species sequence comparison: the next frontier in genome annotation. *Genome Biol*, 4(12):122.

Duehring, U., Axmann, I., Hess, W., and Wilde, A. (2006). An internal antisense rna regulates expression of the photosynthesis gene isia. *submitted*.

Dufresne, A., Garczarek, L., and Partensky, F. (2005). Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol*, 6(2):R14.

Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I., Barbe, V., Duprat, S., Galperin, M., Koonin, E., Le Gall, F., Makarova, K., Ostrowski, M., Oztas, S., Robert, C., Rogozin, I., Scanlan, D., Tandeau de Marsac, N., Weissenbach, J., Wincker, P., Wolf, Y., and Hess, W. (2003). Genome sequence of the cyanobacterium Prochlorococcus marinus SS120, a nearly minimal oxyphototrophic genome. *Proc Natl Acad Sci U S A*, 100(17):10020–5.

Dunn, J. and Studier, F. (1973a). T7 early RNAs and Escherichia coli ribosomal RNAs are cut from large precursor RNAs in vivo by ribonuclease 3. *Proc Natl Acad Sci U S A*, 70(12):3296–3300.

Dunn, J. and Studier, F. (1973b). T7 early RNAs are generated by site-specific cleavages. *Proc Natl Acad Sci U S A*, 70(5):1559–63.

Dunn, J. and Studier, F. (1975). Effect of RNAase III, cleavage on translation of bacteriophage T7 messenger RNAs. *J Mol Biol*, 99(3):487–99.

Dunn, J. and Studier, F. (1983). Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. *J Mol Biol*, 166(4):477–535.

Dvornyk, V., Vinogradova, O., and Nevo, E. (2003). Origin and evolution of circadian clock genes in prokaryotes. *Proc Natl Acad Sci U S A*, 100(5):2495–500.

Ellson, J. and North, S. (2006). Graphviz - Graph Visualization Software. http://www.graphviz.org/.

Ermolaeva, M., Khalak, H., White, O., Smith, H., and Salzberg, S. (2000). Prediction of transcription terminators in bacterial genomes. *J Mol Biol*, 301(1):27–33.

Falkowski, P. (2002). The ocean's invisible forest. *Sci Am*, 287(2):54–61.

Field, C., Behrenfeld, M., Randerson, J., and Falkowski, P. (1998). Primary production of the biosphere: integrating terrestrial and oceanic components. *Science*, 281(5374):237–40.

Franche, C. and Damerval, T. (1988). Tests on nif probes and dna hybridizations. *Methods Enzymol.*, 167:803–8.

Fraser, C. (2006). tigr fams: tigr protein families. http://www.tigr.org/TIGRFAMs/.

Frazer, K., Elnitski, L., Church, D., Dubchak, I., and Hardison, R. (2003). Cross-species sequence comparisons: a review of methods and available resources. *Genome Res*, 13(1):1–12.

Frith, M., Hansen, U., Spouge, J., and Weng, Z. (2004). Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res*, 32(1):189–200.

Fuhrman, J. and Capone, D. (2001). Nifty nanoplankton. *Nature*, 412(6847):593–4.

Fuller, N., Marie, D., Partensky, F., Vaulot, D., Post, A., and Scanlan, D. (2003). Clade-specific 16S ribosomal DNA oligonucleotides reveal the predominance of a single marine Synechococcus clade throughout a stratified water column in the Red Sea. *Appl Environ Microbiol*, 69(5):2430–43.

Garcia-Fernandez, J., Hess, W., Houmard, J., and Partensky, F. (1998). Expression of the psbA gene in the marine oxyphotobacteria Prochlorococcus spp. *Arch Biochem Biophys*, 359(1):17–23.

Garczarek, L., Partensky, F., Irlbacher, H., Holtzendorff, J., Babin, M., Mary, I., Thomas, J., and Hess, W. (2001). Differential expression of antenna and core genes in Prochlorococcus PCC 9511 (Oxyphotobacteria) grown under a modulated light-dark cycle. *Environ Microbiol*, 3(3):168–75.

Gaudin, C., Zhou, X., Williams, K., and Felden, B. (2002). Two-piece tmRNA in cyanobacteria and its structural analysis. *Nucleic Acids Res*, 30(9):2018–24.

Gelfand, M., Koonin, E., and Mironov, A. (2000). Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. *Nucleic Acids Res*, 28(3):695–705.

Gerhart, E., Wagner, H., and Brantl, S. (1998). Kissing and RNA stability in antisense control of plasmid replication. *Trends Biochem Sci*, 23(12):451–4.

Golden, S., Brusslan, J., and Haselkorn, R. (1986). Expression of a family of psbA genes encoding a photosystem II polypeptide in the cyanobacterium Anacystis nidulans R2. *EMBO J*, 5(11):2789–98.

Golden, S. and Canales, S. (2003). Cyanobacterial circadian clocks–timing is everything. *Nat Rev Microbiol*, 1(3):191–9.

Golden, S., Ishiura, M., Johnson, C., and Kondo, T. (1997). CYANOBACTERIAL CIRCADIAN RHYTHMS. *Annu Rev Plant Physiol Plant Mol Biol*, 48:327–354.

Gottesman, S. (1984). Bacterial regulation: global regulatory networks. *Annu Rev Genet*, 18:415–41.

Gottesman, S. (2004). The small RNA regulators of Escherichia coli: roles and mechanisms*. *Annu Rev Microbiol*, 58:303–28.

Gottesman, S. (2005). Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet*, 21(7):399–404.

Gottgens, B., Gilbert, J., Barton, L., Grafham, D., Rogers, J., Bentley, D., and Green, A. (2001). Long-range comparison of human and mouse SCL loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. *Genome Res*, 11(1):87–97.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S., and Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*, 33(Database issue):D121–4.

Grundy, F. and Henkin, T. (2003). The T box and S box transcription termination control systems. *Front Biosci*, 8:d20–31.

Gutekunst, K., Phunpruch, S., Schwarz, C., Schuchardt, S., Schulz-Friedrich, R., and Appel, J. (2005). LexA regulates the bidirectional hydrogenase in the cyanobacterium Synechocystis sp. PCC 6803 as a transcription activator. *Mol Microbiol*, 58(3):810–23.

Hagstrom, A., Pommier, T., Rohwer, F., Simu, K., Stolte, W., Svensson, D., and Zweifel, U. (2002). Use of 16S ribosomal DNA for delineation of marine bacterioplankton species. *Appl Environ Microbiol*, 68(7):3628–33.

Havaux, M., Guedeney, G., Hagemann, M., Yeremenko, N., Matthijs, H., and Jeanjean, R. (2005). The chlorophyll-binding protein IsiA is inducible by high light and protects the cyanobacterium Synechocystis PCC6803 from photooxidative stress. *FEBS Lett*, 579(11):2289–93.

Hawley, D. and McClure, W. (1983). Compilation and analysis of Escherichia coli promoter DNA sequences. *Nucleic Acids Res*, 11(8):2237–55.

Hernandez, J., Muro-Pastor, A., Flores, E., Bes, M., Peleato, M., and Fillat, M. (2006). Identification of a furA cis antisense RNA in the cyanobacterium Anabaena sp. PCC 7120. *J Mol Biol*, 355(3):325–34.

Herrero, A., Muro-Pastor, A., and Flores, E. (2001). Nitrogen control in cyanobacteria. *J Bacteriol*, 183(2):411–25.

Hess, W. (2004). Genome analysis of marine photosynthetic microbes and their global role. *Curr Opin Biotechnol*, 15(3):191–8.

Hess, W. (2006). Cyanolab: HESS lab - research on cyanobacteria. http://www.cyanolab.de/.

Hess, W., Rocap, G., Ting, C., Larimer, F., Stilwagen, S., Lamerdin, J., and Chisholm, S. (2001). The photosynthetic apparatus of Prochlorococcus: Insights through comparative genomics. *Photosynth Res*, 70(1):53–71.

Hu, J., Li, B., and Kihara, D. (2005). Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res*, 33(15):4899–913.

Huntzinger, E., Boisset, S., Saveanu, C., Benito, Y., Geissmann, T., Namane, A., Lina, G., Etienne, J., Ehresmann, B., Ehresmann, C., Jacquier, A., Vandenesch, F., and Romby, P. (2005). Staphylococcus aureus RNAIII and the endoribonuclease III coordinately regulate spa gene expression. *EMBO J*, 24(4):824–35.

Huttenhofer, A., Schattner, P., and Polacek, N. (2005). Non-coding RNAs: hope or hype? *Trends Genet*, 21(5):289–97.

Ihalainen, J., D'Haene, S., Yeremenko, N., van Roon, H., Arteni, A., Boekema, E., van Grondelle, R., Matthijs, H., and Dekker, J. (2005). Aggregates of the chlorophyll-binding protein IsiA (CP43') dissipate energy in cyanobacteria. *Biochemistry*, 44(32):10846–53.

Imamura, S., Asayama, M., and Shirai, M. (2004). In vitro transcription analysis by reconstituted cyanobacterial RNA polymerase: roles of group 1 and 2 sigma factors and a core subunit, RpoC2. *Genes Cells*, 9(12):1175–87.

Imamura, S., Asayama, M., Takahashi, H., Tanaka, K., Takahashi, H., and Shirai, M. (2003a). Antagonistic dark/light-induced SigB/SigD, group 2 sigma factors, expression through redox potential and their roles in cyanobacteria. *FEBS Lett*, 554(3):357–62.

Imamura, S., Yoshihara, S., Nakano, S., Shiozaki, N., Yamada, A., Tanaka, K., Takahashi, H., Asayama, M., and Shirai, M. (2003b). Purification, characterization, and gene expression of all sigma factors of RNA polymerase in a cyanobacterium. *J Mol Biol*, 325(5):857–72.

Imburgio, D., Rong, M., Ma, K., and McAllister, W. (2000). Studies of promoter recognition and start site selection by T7 RNA polymerase using a comprehensive collection of promoter variants. *Biochemistry*, 39(34):10419–30.

Ishihama, A. (2000). Functional modulation of Escherichia coli RNA polymerase. *Annu Rev Microbiol*, 54:499–518.

Ishiura, M., Kutsuna, S., Aoki, S., Iwasaki, H., Andersson, C., Tanabe, A., Golden, S., Johnson, C., and Kondo, T. (1998). Expression of a gene cluster kaiABC as a circadian feedback process in cyanobacteria. *Science*, 281(5382):1519–23.

Ivleva, N., Bramlett, M., Lindahl, P., and Golden, S. (2005). LdpA: a component of the circadian clock senses redox state of the cell. *EMBO J*, 24(6):1202–10.

Iwasaki, H. and Kondo, T. (2004). Circadian timing mechanism in the prokaryotic clock system of cyanobacteria. *J Biol Rhythms*, 19(5):436–44.

Jacquet, S., Partensky, F., Marie, D., Casotti, R., and Vaulot, D. (2001). Cell cycle regulation by light in Prochlorococcus strains. *Appl Environ Microbiol*, 67(2):782–90.

Jeanjean, R., Zuther, E., Yeremenko, N., Havaux, M., Matthijs, H., and Hagemann, M. (2003). A photosystem 1 psaFJ-null mutant of the cyanobacterium Synechocystis PCC 6803 expresses the isiAB operon under iron replete conditions. *FEBS Lett*, 549(1-3):52–6.

Johansson, J. and Cossart, P. (2003). RNA-mediated control of virulence gene expression in bacterial pathogens. *Trends Microbiol*, 11(6):280–5.

Johansson, J., Mandin, P., Renzoni, A., Chiaruttini, C., Springer, M., and Cossart, P. (2002). An RNA thermosensor controls expression of virulence genes in Listeria monocytogenes. *Cell*, 110(5):551–61.

Johnson, C. (2004). Precise circadian clocks in prokaryotic cyanobacteria. *Curr Issues Mol Biol*, 6(2):103–10.

Johnson, P. and Sieburth, J. (1979). Chroococcoid cyanobacteria in the sea - ubiquitous and diverse phototropic biomass. *Limnology and Oceanography*, 24:928–35.

Johnson, Z., Zinser, E., Coe, A., McNulty, N., Woodward, E., and Chisholm, S. (2006). Niche partitioning among prochlorococcus ecotypes along ocean-scale environmental gradients. *Science, in press*.

Kaneko, T., Nakamura, Y., Wolk, C., Kuritz, T., Sasamoto, S., Watanabe, A., Iriguchi, M., Ishikawa, A., Kawashima, K., Kimura, T., Kishida, Y., Kohara, M., Matsumoto, M., Matsuno, A., Muraki, A., Nakazaki, N., Shimpo, S., Sugimoto, M., Takazawa, M., Yamada, M., Yasuda, M., and Tabata, S. (2001). Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium Anabaena sp. strain PCC 7120. *DNA Res*, 8(5):205–13; 227–53.

Kaneko, T. and Tabata, S. (1997). Complete genome structure of the unicellular cyanobacterium Synechocystis sp. PCC6803. *Plant Cell Physiol*, 38(11):1171–6.

Katayama, M., Tsinoremas, N., Kondo, T., and Golden, S. (1999). cpmA, a gene involved in an output pathway of the cyanobacterial circadian system. *J Bacteriol*, 181(11):3516–24.

Kawano, M., Reynolds, A., Miranda-Rios, J., and Storz, G. (2005). Detection of 5'- and 3'-UTR-derived small RNAs and cis-encoded antisense RNAs in Escherichia coli. *Nucleic Acids Res*, 33(3):1040–50.

Keiler, K., Shapiro, L., and Williams, K. (2000). tmRNAs that encode proteolysis-inducing tags are found in all known bacterial genomes: A two-piece tmRNA functions in Caulobacter. *Proc Natl Acad Sci U S A*, 97(14):7778–83.

Kensche, P. (2004). Regulatory rna elements - a motif search in cyanobacterial genomes. Master's thesis, Humboldt University Berlin.

Kensche, P. (2005). Complete results for non-coding rna screening of cyanobacteria. http://itb.biologie.hu-berlin.de/~kensche/ncRNA_05/index.htm.

Kielbasa, S., Gonze, D., and Herzel, H. (2005). Measuring similarities between transcription factor binding sites. *BMC Bioinformatics*, 6:237.

Kielbasa, S., Korbel, J., Beule, D., Schuchhardt, J., and Herzel, H. (2001). Combining frequency and positional information to predict transcription factor binding sites. *Bioinformatics*, 17(11):1019–26.

Kolb, A., Busby, S., Buc, H., Garges, S., and Adhya, S. (1993). Transcriptional regulation by cAMP and its receptor protein. *Annu Rev Biochem*, 62:749–95.

Kondo, T., Mori, T., Lebedeva, N., Aoki, S., Ishiura, M., and Golden, S. (1997). Circadian rhythms in rapidly dividing cyanobacteria. *Science*, 275(5297):224–7.

Korbel, J., Jensen, L., von Mering, C., and Bork, P. (2004). Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat Biotechnol*, 22(7):911–7.

Kramer, R. and Rosenberg, M. (1976). The isolation and characterization of bacteriophage T7 messenger RNA fragments containing an RNase III cleavage site. *Nucleic Acids Res*, 3(10):2411–26.

Kramer, R., Rosenberg, M., and Steitz, J. (1974). Nucleotide sequences of the 5' and 3' termini of bacteriophage T7 early messenger RNAs synthesized in vivo: evidence for sequence specificity in RNA processing. *J Mol Biol*, 89(4):767–76.

Kucho, K., Okamoto, K., Tsuchiya, Y., Nomura, S., Nango, M., Kanehisa, M., and Ishiura, M. (2005). Global analysis of circadian expression in the cyanobacterium Synechocystis sp. strain PCC 6803. *J Bacteriol*, 187(6):2190–9.

Kutsuna, S., Nakahira, Y., Katayama, M., Ishiura, M., and Kondo, T. (2005). Transcriptional regulation of the circadian clock operon kaiBC by upstream regions in cyanobacteria. *Mol Microbiol*, 57(5):1474–84.

La Roche, J., van der Staay, G., Partensky, F., Ducret, A., Aebersold, R., Li, R., Golden, S., Hiller, R., Wrench, P., Larkum, A., and Green, B. (1996). Independent evolution of the prochlorophyte and green plant chlorophyll a/b light-harvesting proteins. *Proc Natl Acad Sci U S A*, 93(26):15244–8.

Lenz, D., Mok, K., Lilley, B., Kulkarni, R., Wingreen, N., and Bassler, B. (2004). The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in Vibrio harveyi and Vibrio cholerae. *Cell*, 118(1):69–82.

Lewin, B. (2000). *genes VII*. Oxford University Press, Oxford University Press Inc., New York.

Lewis, B., Burge, C., and Bartel, D. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20.

Lindahl, L. and Zengel, J. (1986). Ribosomal genes in Escherichia coli. *Annu Rev Genet*, 20:297–326.

Lindell, D., Jaffe, J., Johnson, Z., Church, G., and Chisholm, S. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. *Nature*.

Lindell, D., Sullivan, M., Johnson, Z., Tolonen, A., Rohwer, F., and Chisholm, S. (2004). Transfer of photosynthesis genes to and from Prochlorococcus viruses. *Proc Natl Acad Sci U S A*, 101(30):11013–8.

Logemann, J., Schell, J., and Willmitzer, L. (1987). Improved method for the isolation of RNA from plant tissues. *Anal Biochem*, 163(1):16–20.

Loots, G., Locksley, R., Blankespoor, C., Wang, Z., Miller, W., Rubin, E., and Frazer, K. (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, 288(5463):136–40.

Lopez-Maury, L., Florencio, F., and Reyes, J. (2003). Arsenic sensing and resistance system in the cyanobacterium Synechocystis sp. strain PCC 6803. *J Bacteriol*, 185(18):5363–71.

Madan Babu, M. and Teichmann, S. (2003). Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Res*, 31(4):1234–44.

Mandal, M., Boese, B., Barrick, J., Winkler, W., and Breaker, R. (2003). Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria. *Cell*, 113(5):577–86.

Mandal, M. and Breaker, R. (2004). Gene regulation by riboswitches. *Nat Rev Mol Cell Biol*, 5(6):451–63.

Mann, N. (2003). Phages of the marine cyanobacterial picophytoplankton. *FEMS Microbiol Rev*, 27(1):17–34.

Mann, N., Cook, A., Millard, A., Bailey, S., and Clokie, M. (2003). Marine ecosystems: bacterial photosynthesis genes in a virus. *Nature*, 424(6950):741.

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M., and Penny, D. (2002). Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci U S A*, 99(19):12246–51.

Martinez-Antonio, A. and Collado-Vides, J. (2003). Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol*, 6(5):482–9.

Mazon, G., Erill, I., Campoy, S., Cortes, P., Forano, E., and Barbe, J. (2004a). Reconstruction of the evolutionary history of the LexA-binding sequence. *Microbiology*, 150(Pt 11):3783–95.

Mazon, G., Lucena, J., Campoy, S., Fernandez de Henestrosa, A., Candau, P., and Barbe, J. (2004b). LexA-binding sequences in Gram-positive and cyanobacteria are closely related. *Mol Genet Genomics*, 271(1):40–9.

McDaniel, L., Houchin, L., Williamson, S., and Paul, J. (2002). Lysogeny in marine Synechococcus. *Nature*, 415(6871):496.

McGuire, A., Hughes, J., and Church, G. (2000). Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res*, 10(6):744–57.

Millard, A., Clokie, M., Shub, D., and Mann, N. (2004). Genetic organization of the psbAD region in phages infecting marine Synechococcus strains. *Proc Natl Acad Sci U S A*, 101(30):11007–12.

Molineux, I. (2001). No syringes please, ejection of phage T7 DNA from the virion is enzyme driven. *Mol Microbiol*, 40(1):1–8.

Moller, T., Franch, T., Hojrup, P., Keene, D., Bachinger, H., Brennan, R., and Valentin-Hansen, P. (2002). Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. *Mol Cell*, 9(1):23–30.

Moore, L., Rocap, G., and Chisholm, S. (1998). Physiology and molecular phylogeny of coexisting Prochlorococcus ecotypes. *Nature*, 393(6684):464–7.

Morden, C., Delwiche, C., Kuhsel, M., and Palmer, J. (1992). Gene phylogenies and the endosymbiotic origin of plastids. *Biosystems*, 28(1-3):75–90.

Moreira, D., Le Guyader, H., and Philippe, H. (2000). The origin of red algae and the evolution of chloroplasts. *Nature*, 405(6782):69–72.

Morita, T., Maki, K., and Aiba, H. (2005). RNase E-based ribonucleoprotein complexes: mechanical basis of mRNA destabilization mediated by bacterial noncoding RNAs. *Genes Dev*, 19(18):2176–86.

Mount, D. W. (2001). *Bioinformatics: sequences and genome analysis*. Cold Spring Harbor Laboratority Press, Cold Spring Harbor, New York.

Nahvi, A., Barrick, J., and Breaker, R. (2004). Coenzyme B12 riboswitches are widespread genetic control elements in prokaryotes. *Nucleic Acids Res*, 32(1):143–50.

Nair, U., Ditty, J., Min, H., and Golden, S. (2002). Roles for sigma factors in global circadian regulation of the cyanobacterial genome. *J Bacteriol*, 184(13):3530–8.

Nakajima, M., Imai, K., Ito, H., Nishiwaki, T., Murayama, Y., Iwasaki, H., Oyama, T., and Kondo, T. (2005). Reconstitution of circadian oscillation of cyanobacterial KaiC phosphorylation in vitro. *Science*, 308(5720):414–5.

Nakamura, Y., Kaneko, T., Sato, S., Ikeuchi, M., Katoh, H., Sasamoto, S., Watanabe, A., Iriguchi, M., Kawashima, K., Kimura, T., Kishida, Y., Kiyokawa, C., Kohara, M., Matsumoto, M., Matsuno, A., Nakazaki, N., Shimpo, S., Sugimoto, M., Takeuchi, C., Yamada, M., and Tabata, S. (2002). Complete genome structure of the thermophilic cyanobacterium Thermosynechococcus elongatus BP-1. *DNA Res*, 9(4):123–30.

Onizuka, T., Akiyama, H., Endo, S., Kanai, S., Hirano, M., Tanaka, S., and Miyasaka, H. (2002). CO(2) response element and corresponding trans-acting factor of the promoter for ribulose-1,5-bisphosphate carboxylase/oxygenase genes in Synechococcus sp. PCC7002 found by an improved electrophoretic mobility shift assay. *Plant Cell Physiol*, 43(6):660–7.

Ortmann, A., Lawrence, J., and Suttle, C. (2002). Lysogeny and lytic viral production during a bloom of the cyanobacterium Synechococcus spp. *Microb Ecol*, 43(2):225–31.

Overbeek, R., Begley, T., Butler, R., Choudhuri, J., Chuang, H., Cohoon, M., de Crecy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E., Gerdes, S., Glass, E., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G., Rodionov, D., Ruckert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O., and Vonstein, V. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res*, 33(17):5691–702.

Palenik, B., Brahamsha, B., Larimer, F., Land, M., Hauser, L., Chain, P., Lamerdin, J., Regala, W., Allen, E., McCarren, J., Paulsen, I., Dufresne, A., Partensky, F., Webb, E., and Waterbury, J. (2003). The genome of a motile marine Synechococcus. *Nature*, 424(6952):1037–42.

Palinska, K., Jahns, T., Rippka, R., and Tandeau De Marsac, N. (2000). Prochlorococcus marinus strain PCC 9511, a picoplanktonic cyanobacterium, synthesizes the smallest urease. *Microbiology*, 146 Pt 12:3099–107.

Partensky, F. (2006). Synechococcus sp. wh7803 et rcc307, ubiquitous marine cyanobacteria. http://www.genoscope.cns.fr/externe/English/Projets/Projet_HP/organisme_HP.

Partensky, F., Hess, W., and Vaulot, D. (1999). Prochlorococcus, a marine photosynthetic prokaryote of global significance. *Microbiol Mol Biol Rev*, 63(1):106–27.

Paul, J., Williamson, S., Long, A., Authement, R., John, D., Segall, A., Rohwer, F., Androlewicz, M., and Patterson, S. (2005). Complete genome sequence of phiHSIC, a pseudotemperate marine phage of Listonella pelagia. *Appl Environ Microbiol*, 71(6):3311–20.

Pennacchio, L. and Rubin, E. (2001). Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet*, 2(2):100–9.

Perez-Rueda, E. and Collado-Vides, J. (2000). The repertoire of DNA-binding transcriptional regulators in Escherichia coli K-12. *Nucleic Acids Res*, 28(8):1838–47.

Pichard, S. and Paul, J. (1991). Detection of Gene Expression in Genetically Engineered Microorganisms and Natural Phytoplankton Populations in the Marine Environment by mRNA Analysis. *Appl Environ Microbiol*, 57(6):1721–1727.

Proctor, L. and Fuhrman, J. (1990). Viral mortality of marine bacteria and cyanobacteria. *Nature (London)*, 343:60–2.

Regnier, P. and Grunberg-Manago, M. (1989). Cleavage by RNase III in the transcripts of the met Y-nus-A-infB operon of Escherichia coli releases the tRNA and initiates the decay of the downstream mRNA. *J Mol Biol*, 210(2):293–302.

Repoila, F., Majdalani, N., and Gottesman, S. (2003). Small non-coding RNAs, coordinators of adaptation processes in Escherichia coli: the RpoS paradigm. *Mol Microbiol*, 48(4):855–61.

Reyes, J., Muro-Pastor, M., and Florencio, F. (1997). Transcription of glutamine synthetase genes (glnA and glnN) from the cyanobacterium Synechocystis sp. strain PCC 6803 is differently regulated in response to nitrogen availability. *J Bacteriol*, 179(8):2678–89.

Ripp, S. and Miller, R. (1997). The role of pseudolysogeny in bacteriophage-host interactions in a natural freshwater environment. *Microbiology*, 143:2065–70.

Rippka, R., Deruelles, J., Waterbury, J., Herdman, M., and Stanier, R. (1979). Generic assignments, strain histories and properties of pure cultures of cyanobacteria. *J. Gen. Microbiol.*, 111:1–61.

Rivas, E. and Eddy, S. (2001). Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2:8.

Rivas, E., Klein, R., Jones, T., and Eddy, S. (2001). Computational identification of noncoding RNAs in E. coli by comparative genomics. *Curr Biol*, 11(17):1369–73.

Robertson, H., Dickson, E., and Dunn, J. (1977). A nucleotide sequence from a ribonuclease III processing site in bacteriophage T7 RNA. *Proc Natl Acad Sci U S A*, 74(3):822–6.

Rocap, G., Larimer, F., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N., Arellano, A., Coleman, M., Hauser, L., Hess, W., Johnson, Z., Land, M., Lindell, D., Post, A., Regala, W., Shah, M., Shaw, S., Steglich, C., Sullivan, M., Ting, C., Tolonen, A., Webb, E., Zinser, E., and Chisholm, S. (2003). Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature*, 424(6952):1042–7.

Rojo, F. (1999). Repression of transcription initiation in bacteria. *J Bacteriol*, 181(10):2987–91.

Rosenberg, M. and Kramer, R. (1977). Nucleotide sequence surrounding a ribonuclease III processing site in bacteriophage T7 RNA. *Proc Natl Acad Sci U S A*, 74(3):984–8.

Rosenberg, M., Kramer, R., and Steitz, J. (1974). T7 early messenger RNAs are the direct products of ribonuclease III cleavage. *J Mol Biol*, 89(4):777–82.

Rujan, T. and Martin, W. (2001). How many genes in Arabidopsis come from cyanobacteria? An estimate from 386 protein phylogenies. *Trends Genet*, 17(3):113–20.

Salvador, M., Klein, U., and Bogorad, L. (1998). Endogenous fluctuations of DNA topology in the chloroplast of Chlamydomonas reinhardtii. *Mol Cell Biol*, 18(12):7235–42.

Sambrook, J. and Russell, D. (2001). *Molecular Cloning: A Laboratory Manual, Third Edition*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Schneider, T. and Stephens, R. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20):6097–100.

Schopf, J. (1993). Microfossils of the Early Archean Apex chert: new evidence of the antiquity of life. *Science*, 260:640–6.

Schyns, G., Jia, L., Coursin, T., Tandeau de Marsac, N., and Houmard, J. (1998). Promoter recognition by a cyanobacterial RNA polymerase: in vitro studies with the Calothrix sp. PCC 7601 transcriptional factors RcaA and RcaD. *Plant Mol Biol*, 36(5):649–59.

Shen-Orr, S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet*, 31(1):64–8.

Sledjeski, D., Gupta, A., and Gottesman, S. (1996). The small RNA, DsrA, is essential for the low temperature expression of RpoS during exponential growth in Escherichia coli. *EMBO J*, 15(15):3993–4000.

Steglich, C. (2003). *Biochemical, spectroscopic and moleculargenetic characterisation of novel phycoerythrin species from Prochlorococcus sp.* PhD thesis, Humboldt University Berlin.

Su, Z., Dam, P., Chen, X., Olman, V., Jiang, T., Palenik, B., and Xu, Y. (2003). Computational inference of regulatory pathways in microbes: an application to phosphorus assimilation pathways in Synechococcus sp. WH8102. *Genome Inform Ser Workshop Genome Inform*, 14:3–13.

Su, Z., Olman, V., Mao, F., and Xu, Y. (2005). Comparative genomics analysis of NtcA regulons in cyanobacteria: regulation of nitrogen assimilation and its coupling to photosynthesis. *Nucleic Acids Res*, 33(16):5156–71.

Sullivan, M., Coleman, M., Weigele, P., Rohwer, F., and Chisholm, S. (2005). Three Prochlorococcus Cyanophage Genomes: Signature Features and Ecological Interpretations. *PLoS Biol*, 3(5):e144.

Sullivan, M., Waterbury, J., and Chisholm, S. (2003). Cyanophages infecting the oceanic cyanobacterium Prochlorococcus. *Nature*, 424(6952):1047–51.

Sun, G., Birkey, S., and Hulett, F. (1996). Three two-component signal-transduction systems interact for Pho regulation in Bacillus subtilis. *Mol Microbiol*, 19(5):941–8.

Suttle, C., Chan, A., and Cottrell, M. (1990). Infection of viruses by phytoplankton and reduction of primary productivity. *Nature (London)*, 347:467–9.

Suzuki, S., Ferjani, A., Suzuki, I., and Murata, N. (2004). The SphS-SphR two component system is the exclusive sensor for the induction of gene expression in response to phosphate limitation in synechocystis. *J Biol Chem*, 279(13):13234–40.

Tatusov, R., Fedorova, N., Jackson, J., Jacobs, A., Kiryutin, B., Koonin, E., Krylov, D., Mazumder, R., Mekhedov, S., Nikolskaya, A., Rao, B., Smirnov, S., Sverdlov, A., Vasudevan, S., Wolf, Y., Yin, J., and Natale, D. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41.

Tatusov, R., Koonin, E., and Lipman, D. (1997). A genomic perspective on protein families. *Science*, 278(5338):631–7.

Tegmark, K., Morfeldt, E., and Arvidson, S. (1998). Regulation of agr-dependent virulence genes in Staphylococcus aureus by RNAIII from coagulase-negative staphylococci. *J Bacteriol*, 180(12):3181–6.

Tettelin, H., Masignani, V., Cieslewicz, M., Donati, C., Medini, D., Ward, N., Angiuoli, S., Crabtree, J., Jones, A., Durkin, A., Deboy, R., Davidsen, T., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J., Hauser, C., Sundaram, J., Nelson, W., Madupu, R., Brinkac, L., Dodson, R., Rosovitz, M., Sullivan, S., Daugherty, S., Haft, D., Selengut, J., Gwinn, M., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K., Smith, S., Utterback, T., White, O., Rubens, C., Grandi, G., Madoff, L., Kasper, D., Telford, J., Wessels, M., Rappuoli, R., and Fraser, C. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A*, 102(39):13950–5.

Thiel, T. (1988). Phosphate transport and arsenate resistance in the cyanobacterium Anabaena variabilis. *J Bacteriol*, 170(3):1143–7.

Thompson, J., Higgins, D., and Gibson, T. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–80.

Ting, C., Rocap, G., King, J., and Chisholm, S. (2002). Cyanobacterial photosynthesis in the oceans: the origins and significance of divergent light-harvesting strategies. *Trends Microbiol*, 10(3):134–42.

Tjaden, B., Saxena, R., Stolyar, S., Haynor, D., Kolker, E., and Rosenow, C. (2002). Transcriptome analysis of Escherichia coli using high-density oligonucleotide probe arrays. *Nucleic Acids Res*, 30(17):3732–8.

Tomitani, A., Okada, K., Miyashita, H., Matthijs, H., Ohno, T., and Tanaka, A. (1999). Chlorophyll b and phycobilins in the common ancestor of cyanobacteria and chloroplasts. *Nature*, 400(6740):159–62.

Trotochaud, A. and Wassarman, K. (2004). 6S RNA function enhances long-term cell survival. *J Bacteriol*, 186(15):4978–85.

Trotochaud, A. and Wassarman, K. (2005). A highly conserved 6S RNA structure is required for regulation of transcription. *Nat Struct Mol Biol*, 12(4):313–9.

Urbach, E., Robertson, D., and Chisholm, S. (1992). Multiple evolutionary origins of prochlorophytes within the cyanobacterial radiation. *Nature*, 355(6357):267–70.

Ureta-Vidal, A., Ettwiller, L., and Birney, E. (2003). Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet*, 4(4):251–62.

Valentin-Hansen, P., Eriksen, M., and Udesen, C. (2004). The bacterial Sm-like protein Hfq: a key player in RNA transactions. *Mol Microbiol*, 51(6):1525–33.

Vazquez-Bermudez, M., Flores, E., and Herrero, A. (2002). Analysis of binding sites for the nitrogen-control transcription factor NtcA in the promoters of Synechococcus nitrogen-regulated genes. *Biochim Biophys Acta*, 1578(1-3):95–8.

Venter, J., Remington, K., Heidelberg, J., Halpern, A., Rusch, D., Eisen, J., Wu, D., Paulsen, I., Nelson, K., Nelson, W., Fouts, D., Levy, S., Knap, A., Lomas, M., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y., and Smith, H. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304(5667):66–74.

Verhamme, D., Arents, J., Postma, P., Crielaard, W., and Hellingwerf, K. (2002). Investigation of in vivo cross-talk between key two-component systems of Escherichia coli. *Microbiology*, 148(Pt 1):69–78.

Vinnemeier, J., Kunert, A., and Hagemann, M. (1998). Transcriptional analysis of the isiAB operon in salt-stressed cells of the cyanobacterium Synechocystis sp. PCC 6803. *FEMS Microbiol Lett*, 169(2):323–30.

Vogel, J., Argaman, L., Wagner, E., and Altuvia, S. (2004). The small RNA IstR inhibits synthesis of an SOS-induced toxic peptide. *Curr Biol*, 14(24):2271–6.

Vogel, J., Axmann, I., Herzel, H., and Hess, W. (2003a). Experimental and computational analysis of transcriptional start sites in the cyanobacterium Prochlorococcus MED4. *Nucleic Acids Res*, 31(11):2890–9.

Vogel, J., Bartels, V., Tang, T., Churakov, G., Slagter-Jager, J., Huttenhofer, A., and Wagner, E. (2003b). RNomics in Escherichia coli detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res*, 31(22):6435–43.

Vogel, J. and Sharma, C. (2005). How to find small non-coding RNAs in bacteria. *Biol Chem*, 386(12):1219–38.

Wagner, E. and Simons, R. (1994). Antisense RNA control in bacteria, phages, and plasmids. *Annu Rev Microbiol*, 48:713–42.

Walker, G. (1984). Mutagenesis and inducible responses to deoxyribonucleic acid damage in Escherichia coli. *Microbiol Rev*, 48(1):60–93.

Washietl, S. and Hofacker, I. (2004). Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J Mol Biol*, 342(1):19–30.

Washietl, S., Hofacker, I., and Stadler, P. (2005). Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A*, 102(7):2454–9.

Wassarman, K., Repoila, F., Rosenow, C., Storz, G., and Gottesman, S. (2001). Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev*, 15(13):1637–51.

Wassarman, K. and Storz, G. (2000). 6S RNA regulates E. coli RNA polymerase activity. *Cell*, 101(6):613–23.

Wasserman, W. and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet*, 5(4):276–87.

Watanabe, T., Sugiura, M., and Sugita, M. (1997). A novel small stable RNA, 6Sa RNA, from the cyanobacterium Synechococcus sp. strain PCC6301. *FEBS Lett*, 416(3):302–6.

Waterbury, J., Watson, S., Guillard, R., and Brand, L. (1979). Widespread occurrence of a unicellular, marine, planktonic, cyanobacterium. *Nature*, 277(5694):293–4.

Waterbury, J., Watson, S., Valois, F., and Franks, D. (1986). Biologial and ecological characterization of the marine unicellular cyanobacterium synechococcus. *T. Platt and W. Li (eds.), Photosynthetic Picoplankton. Can. J. Fish. Aquat. Sci. Bull.*, 214:71–120.

Waterbury, J. and Willey, J. (1988). Isolation and growth of marine planktonic cyanobacteria. *Methods in Enzymology*, 167:100–5.

Webb, R., Reddy, K., and Sherman, L. (1990). Regulation and sequence of the Synechococcus sp. strain PCC 7942 groESL operon, encoding a cyanobacterial chaperonin. *J Bacteriol*, 172(9):5079–88.

Weilbacher, T., Suzuki, K., Dubey, A., Wang, X., Gudapaty, S., Morozov, I., Baker, C., Georgellis, D., Babitzke, P., and Romeo, T. (2003). A novel sRNA component of the carbon storage regulatory system of Escherichia coli. *Mol Microbiol*, 48(3):657–70.

West, N. and Scanlan, D. (1999). Niche-partitioning of Prochlorococcus populations in a stratified water column in the eastern North Atlantic Ocean. *Appl Environ Microbiol*, 65(6):2585–91.

Wheeler, D., Barrett, T., Benson, D., Bryant, S., Canese, K., Church, D., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W., Kenton, D., Khovayko, O., Lipman, D., Madden, T., Maglott, D., Ostell, J., Pontius, J., Pruitt, K., Schuler, G., Schriml, L., Sequeira, E., Sherry, S., Sirotkin, K., Starchenko, G., Suzek, T., Tatusov, R., Tatusova, T., Wagner, L., and Yaschenko, E. (2005). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 33(Database issue):D39–45.

Wilderman, P., Sowa, N., FitzGerald, D., FitzGerald, P., Gottesman, S., Ochsner, U., and Vasil, M. (2004). Identification of tandem duplicate regulatory small RNAs in Pseudomonas aeruginosa involved in iron homeostasis. *Proc Natl Acad Sci U S A*, 101(26):9792–7.

Williams, K. (2002). Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res*, 30(4):866–75.

Wilson, W., Carr, N., and Mann, N. (1996). The effect of phosphate status on the kinetics of cyanophage infection of the oceanic cyanobacterium synechococcus sp. wh7803. *Journal of Phycology*, 32:506–16.

Winkler, W., Nahvi, A., and Breaker, R. (2002). Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature*, 419(6910):952–6.

Xie, W., Jager, K., and Potts, M. (1989). Cyanobacterial RNA polymerase genes rpoC1 and rpoC2 correspond to rpoC of Escherichia coli. *J Bacteriol*, 171(4):1967–73.

Xu, Y., Mori, T., and Johnson, C. (2003). Cyanobacterial circadian clockwork: roles of KaiA, KaiB and the kaiBC promoter in regulating KaiC. *EMBO J*, 22(9):2117–26.

Yan, B., Methe, B., Lovley, D., and Krushkal, J. (2004). Computational prediction of conserved operons and phylogenetic footprinting of transcription regulatory elements in the metal-reducing bacterial family Geobacteraceae. *J Theor Biol*, 230(1):133–44.

Young, M. and Kay, S. (2001). Time zones: a comparative genetics of circadian clocks. *Nat Rev Genet*, 2(9):702–15.

Yousef, N., Pistorius, E., and Michel, K. (2003). Comparative analysis of idiA and isiA transcription under iron starvation and oxidative stress in Synechococcus elongatus PCC 7942 wild-type and selected mutants. *Arch Microbiol*, 180(6):471–83.

Zago, M., Dennis, P., and Omer, A. (2005). The expanding world of small RNAs in the hyperthermophilic archaeon Sulfolobus solfataricus. *Mol Microbiol*, 55(6):1812–28.

Zengel, J. and Lindahl, L. (1994). Diverse mechanisms for regulating ribosomal protein synthesis in Escherichia coli. *Prog Nucleic Acid Res Mol Biol*, 47:331–70.

Zhang, A., Altuvia, S., Tiwari, A., Argaman, L., Hengge-Aronis, R., and Storz, G. (1998). The OxyS regulatory RNA represses rpoS translation and binds the Hfq (HF-I) protein. *EMBO J*, 17(20):6061–8.

Zhang, A., Wassarman, K., Ortega, J., Steven, A., and Storz, G. (2002). The Sm-like Hfq protein increases OxyS RNA interaction with target mRNAs. *Mol Cell*, 9(1):11–22.

Zhang, A., Wassarman, K., Rosenow, C., Tjaden, B., Storz, G., and Gottesman, S. (2003). Global analysis of small RNA and mRNA targets of Hfq. *Mol Microbiol*, 50(4):1111–24.

Zhang, K. and Nicholson, A. (1997). Regulation of ribonuclease III processing by double-helical sequence antideterminants. *Proc Natl Acad Sci U S A*, 94(25):13437–41.

Zhang, X. and Studier, F. (1997). Mechanism of inhibition of bacteriophage T7 RNA polymerase by T7 lysozyme. *J Mol Biol*, 269(1):10–27.

Zhang, X. and Studier, F. (2004). Multiple roles of T7 RNA polymerase and T7 lysozyme during bacteriophage T7 infection. *J Mol Biol*, 340(4):707–30.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, 31(13):3406–15.

Appendix

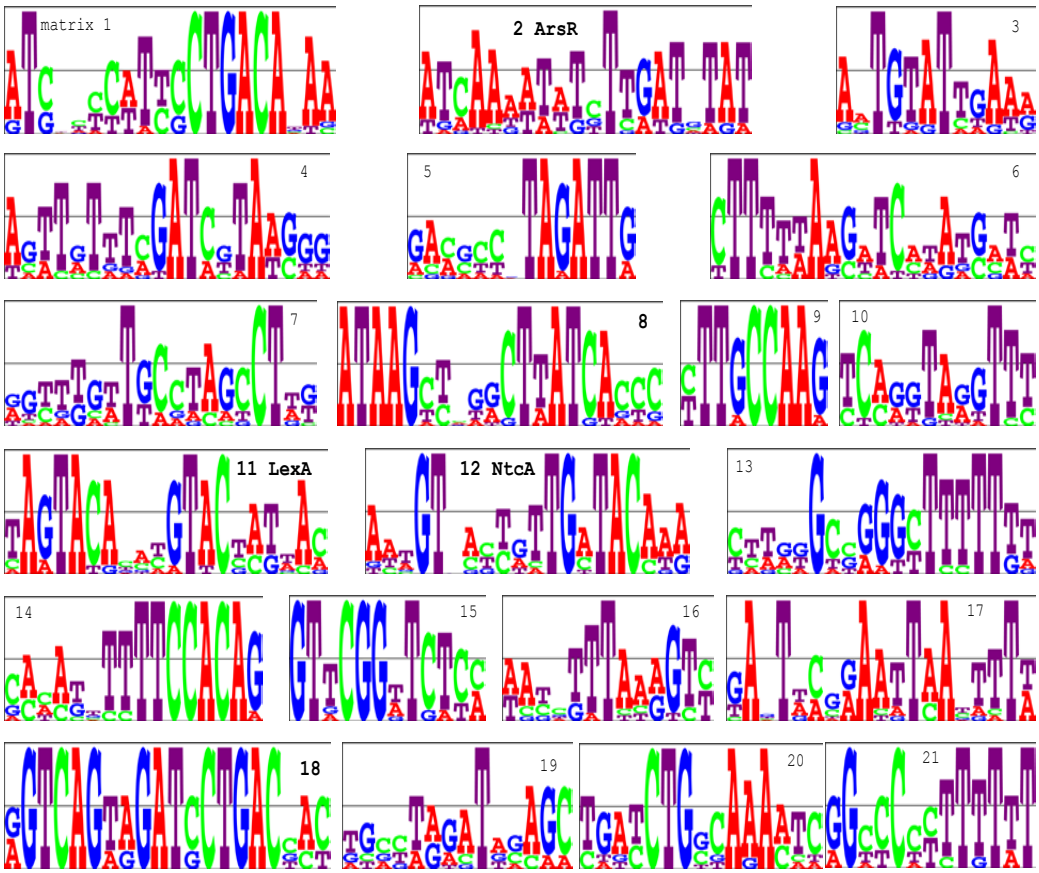


Figure 2: Sequence logos of the best conserved elements obtained by phylogenetic footprinting analysis of *Prochlorococcus* Med4, MIT 9313, SS120 and *Synechococcus* WH 8102. Based on these 21 clusters, position-specific weight matrices (PSWM) were constructed to search for candidate regulatory elements in upstream regions of all four genomes. Three predicted clusters bear analogy to already known sites for certain cyanobacteria: ArsR (matrix 2), LexA (matrix 11) and NtcA (matrix 12). The focus on spaced and palindromic DNA motifs revealed two additional best clusters (matrix 8 and 18).

Med4	MIT 9313	SS120	WH 8102	function
PMM1697	PMT2246	Pro1863	SYNW2496	Type II alternative RNA polymerase sigma factor
PMM1637	PMT0118	Pro1799	SYNW0091	cyanobacteria-specific GntR-like HTH domain containing transcriptional regulator
PMM1629	PMT0127	Pro1791	SYNW0102	Type II alternative RNA polymerase sigma factor
PMM1619	PMT0146	Pro1780	SYNW0126	Two-component response regulator
PMM1494	PMT1528	Pro1649	SYNW0597	Translation initiation factor 2
PMM1490	PMT1512	Pro1645	SYNW0608	Trypsin-like serine protease
PMM1463	PMT1481	Pro1616	SYNW0462	Nitrogen regulatory protein PII
PMM1444	PMT1459	Pro1597	SYNW0504	Membrane associated GTPase
PMM1341	PMT1417	Pro1422	SYNW0551	Sensory protein (HAMP domain e: 9.5e-17)
PMM1289	PMT0346	Pro1363	SYNW1621	Type II alternative RNA polymerase sigma factor
PMM1281	PMT0355	Pro1355	SYNW1613	GTPase, Era homolog
PMM1262	PMT0380	Pro1336	SYNW1582	SOS-response transcriptional repressor LexA
PMM1138	PMT1155	Pro1232	SYNW0704	Predicted GTPase, probable translation factor
PMM1113	PMT1097	Pro1083	SYNW0904	Two-component response regulator
PMM1030	PMT2208	Pro1502	SYNW2401	Fe2+/Zn2+ uptake regulation protein
PMM0916	PMT0597	Pro1033	SYNW1078	Membrane associated GTPase
PMM0762	PMT0575	Pro0834	SYNW1109	Predicted membrane GTPase
PMM0714	PMT1001	Pro1578	SYNW0798	Transcriptional regulator ArsR-like
PMM0637	PMT0858	Pro0794	SYNW0865	Fe2+/Zn2+ uptake regulation protein
PMM0577	PMT0456	Pro0580	SYNW1509	Type II alternative RNA polymerase sigma factor
PMM0573	PMT0464	Pro0575	SYNW1518	AbrB family transciptional regulator
PMM0515	PMT1252	Pro0515	SYNW1763	Phosphoglycerate mutase
PMM0496	PMT1272	Pro0495	SYNW1783	Putative principal RNA polymerase sigma factor (sigA,rpoD)
PMM0482	PMT0908	Pro0481	SYNW1909	Membrane protease subunits
PMM0443	PMT1344	Pro0439	SYNW1864	AbrB family transciptional regulator
PMM0391	PMT0203	Pro0388	SYNW0657	Predicted GTPase
PMM0262	PMT1852	Pro0294	SYNW0254	DNA-binding protein, stimulates sugar fermentation
PMM0246	PMT1831	Pro0277	SYNW0275	Global nitrogen regulator NtcA
PMM0189	PMT2074	Pro0214	SYNW2322	Predicted GTPase
PMM0169	PMT2043	Pro0194	SYNW2289	Two-component response regulator
PMM0155	PMT2015	Pro0179	SYNW2269	Uncharacterized conserved protein (CYTH domain e: 7.6e-05)
PMM0154	PMT2014	Pro0177	SYNW2268	Transcriptional regulator, contains HTH motif (luxR family e: 3.2e-21)
PMM0152	PMT2012	Pro0175	SYNW2266	Nucleoside-diphosphate-sugar transferase
PMM0147	PMT2007	Pro0170	SYNW2260	RuBisCO operon transcriptional regulator
PMM0134	PMT1994	Pro0156	SYNW2246	Two-component response regulator (cztR_silR_copR: heavy metal response e: 1e-61)
PMM0128	PMT1988	Pro0150	SYNW2236	Two-component response regulator
PMM0090	PMT1636	Pro0106	SYNW2176	Periplasmic trypsin-like serine protease
PMM0080	PMT1623	Pro0094	SYNW2163	Predicted transcriptional regulator
PMM0077	PMT1618	Pro0091	SYNW2158	ATPase

Table 1: A core set of 35 regulatory factor and 5 σ factor genes conserved between *Prochlorococcus* Med4, MIT 9313, SS120 and *Synechococcus* WH 8102 was revealed by 4-way BLASTp comparison with a cutoff E-value of 10^{-5} and comparison to The SEED database (Overbeek et al., 2005), TIGRFAMs/Pfams "Regulatory functions" (Fraser, 2006). Putative functions were assigned from the SS120 genome annotation (and from TIGRFAM hits to SS120).

downstream gene	position	score	hit sequence
1. motif; matrix 2		(cutoff score: 15)	ArsR-like
PMM0010	49	17.71367	ATAAATCAATAATTTTTTAT
PMM0214	71	15.59813	ATTTTTTAACAAATTTTGAT
PMM0215	176	16.56511	ATCAAAATTTGTTAAAAAAT
PMM0335	293	15.14369	TTAAATTTATTTTAATCTGA
PMM0359	216	17.32203	ATAAAAAATTTCTGATTTAT
PMM0414	24	16.06529	ATAGATTAAGTAATTTTTAA
PMM0465	28	15.35478	CTATAATAAGAAAAATTGAT
PMM0482	247	16.65583	AGAAATCGAGAAATTTTGTT
PMM0506	54	15.74686	TTCAATAATTTTGAATATAT
PMM0519	24	15.53282	ATAAAAATATATTGTTATGT
PMM0619	65	17.48895	ATACATTATGATTTTTTGAA
PMM0622	83	18.6638	TTCAAAATTTTTTAATTTAA
PMM0623	35	18.39738	TTAAATTAAAAAATTTTGAA
PMM0705	49	15.88768	TTACAACAAATAAATTTGAT
PMM0709	143	17.45089	TTAAAAATATGTCAATCAAT
PMM0710	235	17.95111	ATTGATTGACATATTTTTAA
PMM0713	49	27.79706	ATACATCAAAATATTTTGAT
PMM0713	45	27.79287	ATCAAAATATTTTGATGTAT
PMM0714	31	27.69488	ATACATCAAAATATTTTGAT
PMM0714	27	27.69031	ATCAAAATATTTTGATGTAT
PMM0717	137	19.09974	ATAAATTAAAAAATATTTAA
PMM0718	64	18.21427	TTAAATATTTTTTAATTTAT
PMM0726	123	15.0388	TTAAATTAAGAAATTTGTAT
PMM0947	58	16.95098	ATTAATCAAAATTACTTGAA
PMM0975	250	15.05821	ATAAATTAAATTTGATGAAT
PMM1077	62	19.38194	ATATAAATATCTTAATGTGT
PMM1078	64	19.37619	ACACATTAAGATATTTATAT
PMM1118	252	15.16239	ATACCTGAAAAAATTTTTTT
PMM1155	94	15.41322	ATTTTTTAAGATATTTTTAT
PMM1197	156	17.5173	NTAAATCAAAATTAATTTAT
medRNA_44	24	16.58527	ATCAAAAATTGCTAATATAT
PMM1310	41	16.66728	ATATATTAGCAATTTTTGAT
PMM1404	277	16.39646	TTATCTTAAAAAATTTTTAA
PMM1413	59	15.92726	ATCAATTATTTTTGATTGAT
PMM1440	77	15.73466	AGCAAGAATTTTTGATCAAA
PMM1511	72	15.38256	ATCAAAAATGCTTTAGATAT
PMM1512	142	15.81907	ATATCTAAAGCATTTTTGAT
PMM1588	125	18.52823	ATAAAATTTTTTTGAGAAAT
PMM1707	291	17.16661	TTAAAAAAATCTTTATTTAT
PMM1708	20	15.24595	ATAAATAAAGATTTTTTTAA
PMT0282	292	18.04943	TGCTCAATTTCTTGATGTAT
PMT0283	236	17.97373	ATACATCAAGAAATTGAGCA
PMT0304	273	20.0033	ATACCTCGAGCAATTTTCAT
PMT0305	26	19.55711	ATGAAAATTGCTCGAGGTAT
PMT0993	94	24.75818	ATATATCAAGATTTCTTGAT
PMT0993	90	21.98635	ATCAAGATTTCTTGATGGAT
PMT1000	28	29.12228	ATACATCAAGATATTTTGAT
PMT1000	24	18.5389	ATCAAGATATTTTGATTGAT
PMT1001	98	18.83245	ATCAATCAAAATATCTTGAT
PMT1001	94	29.41211	ATCAAAATATCTTGATGTAT
PMT1045	138	16.77599	ATCAAGATATTCTGATAAAT
PMT1435	112	15.73631	ATACGTCAATAGATTTTGAT
Pro0259	129	20.8185	ATAGAGCAACAAAATTTGAT
Pro0260	142	20.83304	ATCAAATTTTGTTGCTCTAT
Pro0263	40	21.583	TTCAATATATTTTGATGAAT
Pro0264	38	21.58681	ATTCATCAAAATATATTGAA
Pro0378	41	15.42825	ATAAAATAAAAAAACATGAT
Pro0510	293	15.61229	ATAAATTATTCTTGTTAAGT
Pro1002	162	15.09725	ATAAATCCAGAGATTTTCTT
Pro1379	76	15.47173	ATATCACAAGCAAAAGTGAT
Pro1494	152	16.89803	AAAAAAATATTTTAATTTGT
Pro1559	29	15.89025	ATAGATTAAACAATCTTTTT
Pro1572	79	16.47627	ATACATAAAGAGTATTTTAT
Pro1573	95	16.51243	ATAAAATACTCTTTATGTAT
Pro1574	57	15.76892	TTATATCAAGTATTATTGAT
Pro1575	72	26.76199	ATACATCAAGAAAACTTGAT
Pro1577	55	18.7366	ATATATCAATAATTCTTGAT
Pro1577	51	24.30857	ATCAATAATTCTTGATATAT
Pro1578	88	24.5432	ATATATCAAGAATTATTGAT
Pro1578	84	18.92793	ATCAAGAATTATTGATATAT

Pro1642	74	15.39444	TTTGAACAAGATATTTTTTT
Pro1643	79	15.45753	AAAAAAATATCTTGTTCAAA
Pro1667	252	15.07919	AGAAATCAAGCTTTTTAGAT
Pro1847	110	16.17289	ATTAAACAAGAATATTTGTT
Pro1847	103	16.19721	AAGAATATTTGTTGATCTAA
ss1RNA_44	239	16.96463	TTAGATCAACAAATATTCTT
ss1RNA_44	232	16.98786	AACAAATATTCTTGTTTAAT
SYNW0798	46	23.51881	ATCAAATTATGTTGATCTAA
SYNW0799	37	23.47682	TTAGATCAACATAATTTGAT
wh8RNA_40	163	15.6728	ATAGATAGAGAAATCTAGAT
wh8RNA_53	163	15.6728	ATAGATAGAGAAATCTAGAT
2. motif; matrix 8		(cutoff score: 10)	
PMM0194	33	10.8534	GGTTCATACGTACAACTTGT
PMM0195	212	11.11291	ACAAGTTGTACGTATGAACC
PMM0344	146	12.45796	ATAAATTTTGGTTATCAGTA
PMM0548	273	12.41028	TACTGATAAGTCAAGCATTT
PMM0548	221	21.16172	CATTGATAAGTTAAACTTAT
PMM0548	216	27.54579	ATAAGTTAAACTTATCACCC
PMM0549	145	27.32375	GGGTGATAAGTTTAACTTAT
PMM0549	140	20.74804	ATAAGTTTAACTTATCAATG
PMM0549	88	11.84022	AAATGCTTGACTTATCAGTA
PMM1352	131	12.25501	ATAAGTTTTTCTAATGTAAC
PMM1378	39	16.12585	CTAAGCTATGCTAATCAAAG
PMM1485	40	10.43569	TGGTAATTAGCAAATCTTAC
PMT0400	221	10.04086	GACTCATCAATATGACTTAT
PMT0423	216	14.99887	AAAAGCTTGGCTAATCATTG
PMT0544	169	12.43199	ATAAGACTTACTAATCCATG
PMT0847	37	12.03551	TAGTGATAACTTCGACTAAT
PMT0870	49	10.00709	ATCAATCAGACTTATCATTC
PMT1206	130	30.34871	GGGTGATTAGCCGAGCTTAT
PMT1207	226	30.29095	ATAAGCTCGGCTAATCACCC
PMT1250	23	12.16054	GATTTATTAGAGTGGCTTAT
PMT1387	230	12.99374	GGCTGATTAGCTGGAATTTT
PMT1429	146	20.31483	ATAAGCCAGAGTTATCAGTA
PMT2077	270	11.5046	CTCTGATTAATCTAGCTTGT
Pro0056	251	13.15664	GATTCATAAGAAAAATTTAT
Pro0057	208	13.03005	ATAAATTTTTCTTATGAATC
Pro0146	21	12.22789	CGGTGAAAAACTAATCTTAT
Pro0296	71	11.10765	TGGCGTTAAGCCAAGCTTTT
Pro0431	69	10.8438	TAGTCATAACTAAGGCATAT
Pro0432	139	11.0982	ATATGCCTTAGTTATGACTA
Pro0549	252	21.06009	CGCTGATAAGCTGTGCTTAT
Pro0549	247	29.62511	ATAAGCTGTGCTTATCACCC
Pro0550	132	29.59745	GGGTGATAAGCACAGCTTAT
Pro0550	127	21.0197	ATAAGCACAGCTTATCAGCG
Pro0715	92	11.81753	ATAATCCAGGCTTATCTTTA
Pro0718	28	10.32374	GTTTGATTAATATGGCTTAA
Pro1146	58	13.87713	ATAAGCTCTTGGTATCCCCC
Pro1434	159	21.33937	CATTGATAAGTCAATCTTAT
Pro1443	130	10.0577	GCGTAATAAAACAAACTTAT
Pro1492	245	12.43014	TAGTAATTTGTATAGCTTAT
Pro1527	174	19.2668	GACAGATCAGCCTGGCTTAT
Pro1527	169	10.41246	ATCAGCCTGGCTTATGCATG
Pro1528	121	10.47057	CATGCATAAGCCAGGCTGAT
Pro1528	116	19.3178	ATAAGCCAGGCTGATCTGTC
Pro1579	157	10.19982	NTGAGATAAATAGTGCTTAT
SYNW0403	47	14.68371	AGAAGTTGGTCTAATCTGCC
SYNW0404	49	14.667	GGCAGATTAGACCAACTTCT
SYNW0535	150	12.09009	AAGCAATAAGCCAGGCTTAT
SYNW0535	145	25.45269	ATAAGCCAGGCTTATCAGTG
SYNW0610	51	10.27696	TGGTGATCAGTCCGGGCTAT
SYNW0611	64	10.20396	ATAGCCCGGACTGATCACCA
SYNW0983	150	14.29245	AGCTGATAAGCGGAGCCTAT
SYNW1179	119	10.67813	GGGTTATAAGCAAAAAATAT
SYNW1719	129	31.11245	GGGTGATAAGCCGGGCTTAT
SYNW1719	124	17.86779	ATAAGCCGGGCTTATCAGGT
SYNW1720	246	17.79714	ACCTGATAAGCCCGGCTTAT
SYNW1720	241	30.94115	ATAAGCCCGGCTTATCACCC
SYNW2224	295	10.77148	ATAAGCTCAGGATCTCAACA
3. motif; matrix 11		(cutoff score: 10)	LexA
PMM0273	22	13.01708	AAATATACATGGACTATTAG
PMM0334	268	10.01359	CTAGTCGTACTGATGTCCTT

PMM0334	36	21.64735	TTTATAGTATATTTGTACTA
PMM0334	32	19.23182	TAGTATATTTGTACTATCAA
PMM0337	290	20.5413	CTAATAATACATCTGTACTA
PMM0337	286	19.82232	TAATACATCTGTACTAATAA
PMM0337	266	12.27045	TAGTACAAAAATACCGTTGC
PMM0383	196	11.87095	TTCATAATACAAATATACTT
PMM0779	58	15.15134	TTAAGAGTACTTTTTTATTA
PMM0819	27	17.56949	TTAGTAGTACACATGTATTA
PMM0819	23	17.09537	TAGTACACATGTATTACTAA
PMM0936	60	17.83723	TATATAGTATATATGTACTA
PMM0936	56	23.47661	TAGTATATATGTACTATTAA
PMM1134	87	14.10629	TAGTAAATGAGTACTTTAAC
PMM1134	28	14.74932	CTAATAGTATAAATGTACCA
PMM1134	24	21.2969	TAGTATAAATGTACCATTAA
PMM1388	123	14.36985	GTATTAGTATTGATGTATTG
PMM1389	25	13.90304	CAATACATCAATACTAATAC
PMM1427	27	14.49208	ATATTAGTACACTTGTACTA
PMM1427	23	13.12711	TAGTACACTTGTACTAACTA
PMM1562	100	15.85654	GCTAAAGTACACATGTACTA
PMM1562	96	23.38556	AAGTACACATGTACTATGAA
PMT0090	41	12.07384	CAGTACAGCTGTAGCTGTAA
PMT0118	160	11.56402	CAGTATGCAAATTCTCTAAA
PMT0221	49	10.90696	CACTACTTACGTACTACAAC
PMT0222	113	11.09087	GTTGTAGTACGTAAGTAGTG
PMT0302	169	20.85696	CAATACACCTGTACTAGTCC
PMT0303	22	20.92371	GGACTAGTACAGGTGTATTG
PMT0635	25	19.57433	CTGCTGGAACGGGTGTACTA
PMT0637	38	13.3055	TAGTACAAAGATACTAGTGC
PMT0638	57	13.26128	GCACTAGTATCTTTGTACTA
PMT0834	34	10.71573	GCTGTAGAACATGAGTACTA
PMT0841	34	15.01391	TTGATAGAACATAAGTACTG
PMT1009	31	20.32052	GGGCTAGTACAGGTGTACTT
mitRNA_37	22	18.19112	GTGATGGTTCTGCTGTACTA
PMT1398	133	18.20555	TAGTACAGCAGAACCATCAC
PMT1661	88	11.01718	TAATTAGTTCATATGTACTA
PMT1661	84	20.0101	TAGTTCATATGTACTATTCC
PMT1727	150	25.96999	GTTAGGGTACGTCTGTACTA
PMT2115	23	17.68274	GAGCTAGTACAGGTGTACTT
PMT2116	282	14.68024	NGTCAAGTACACCTGTACTA
PMT2116	278	17.69666	AAGTACACCTGTACTAGCTC
PMT2209	274	10.09333	CAGTTCATCCATTCTCTCAA
Pro0031	35	13.26159	NNNNGAGTACAAATGTATTA
Pro0623	114	22.90965	GTAATAGTATAAATGTACTA
Pro0623	110	13.86032	TAGTATAAATGTACTACTGC
Pro0624	84	13.77464	GCAGTAGTACATTTATACTA
Pro0624	80	22.65136	TAGTACATTTATACTATTAC
Pro0685	33	12.48103	AGACTAGTATACTTGTACTA
Pro0685	29	18.67888	TAGTATACTTGTACTACTAC
Pro0686	179	19.28547	GTAGTAGTACAAGTATACTA
Pro0686	175	13.00231	TAGTACAAGTATACTAGTCT
Pro0703	249	14.60367	TAACTAGTACTTAAGTACTA
Pro0703	245	13.77225	TAGTACTTAAGTACTAGCCG
Pro0704	38	13.50878	CGGCTAGTACTTAAGTACTA
Pro0704	34	13.72053	TAGTACTTAAGTACTAGTTA
Pro0709	52	10.03042	TTAATGATATAAGTTTACTA
Pro0710	121	10.47232	TAGTAAACTTATATCATTAA
Pro0760	37	25.8413	GTTATGGTACATTTGTACTA
Pro0760	33	10.96035	TGGTACATTTGTACTATCTG
Pro0914	184	12.59865	TTTATAATATAATTGTATTA
Pro0915	23	11.11338	TAATACAATTATATTATAAA
Pro1231	97	11.73546	TTTCTAGAATATCTATATTG
Pro1247	38	11.73832	TAGTTTGTTTGTACTATTAG
Pro1336	48	10.7095	CTTTAGGTACATATGTATTG
Pro1484	131	11.05655	TCTATAGTACATAGTTACTA
Pro1484	127	12.01076	TAGTACATAGTTACTATAGA
Pro1485	48	11.60802	TCTATAGTAACTATGTACTA
Pro1485	44	10.67737	TAGTAACTATGTACTATAGA
Pro1716	108	26.65132	GTTAGCGTACGTGTGTACTA
Pro1839	27	11.30375	TAGTAGCGCTGTTCTCTTAA
Pro1840	119	11.46455	TTAAGAGAACAGCGCTACTA
SYNW0037	20	14.18618	GAACTGATACAGGTGTACTG
SYNW1044	40	20.90725	CAATACATCAGTACTATTAA

SYNW1045	71	21.07099	TTAATAGTACTGATGTATTG
SYNW1138	31	22.37538	GTACTAGTACTGGTGTATTG
SYNW1140	32	24.14263	GTACTAGTACTTGCGTACTA
SYNW1140	28	10.41996	TAGTACTTGCGTACTAGCGG
SYNW1405	21	12.84859	GAACCAGTATGGCTGTACTA
SYNW1466	20	11.12337	GCTTCAGTACTGATGTACTA
SYNW1467	39	11.76793	NNNNTAGTACATCAGTACTG
SYNW1467	35	11.17171	TAGTACATCAGTACTGAAGC
SYNW1660	192	11.04585	TAATACACCTGTACCAACTG
SYNW2062	83	23.34863	GTTAGCGTACGCCTGTACTA
SYNW2104	80	12.71489	TGGTACAGATATATTAGTAC
SYNW2105	198	13.15952	GTACTAATATATCTGTACCA
SYNW2106	33	15.97836	TGTTTAGAACAGGTGTACTA
SYNW2106	29	15.69584	TAGAACAGGTGTACTACTAC
4. motif; matrix 12		(cutoff score: 12)	NtcA
PMM0120	56	13.85507	ATAATTACAATTGATACAAA
PMM0121	107	14.08338	TTTGTATCAATTGTAATTAT
PMM0245	159	26.91616	CTTGTATCAACAGTAACATT
PMM0246	65	26.58152	AATGTTACTGTTGATACAAG
PMM0378	128	12.93099	AACCTCTTTTTTGATACAAA
PMM0404	211	13.27921	GTGGTTACTTTTGATATCAA
PMM0417	181	16.56816	AATGTAGCAATAGCTACTTT
PMM0919	167	23.79687	TATGTATCAAAGGTAACTTT
PMM0920	67	23.27283	AAAGTTACCTTTGATACATA
PMM0970	63	21.58554	AATGTTACCTATGCTACAAA
PMM1118	131	12.97309	TATGTATCAAATATGTCTTT
PMM1164	98	12.30187	TATGTATTAATTGATACAAG
PMM1461	132	24.79925	ATTGTTATCATTGATACAAA
PMM1462	67	24.38745	TTTGTATCAATGATAACAAT
PMT0509	26	16.7513	ATGGTTGGCGTTGATACCTG
PMT0580	255	12.72427	GATGTGCTTTCTGCTACATC
PMT0601	81	23.84146	AAGGTACCTGTTGCTACAAA
PMT1023	41	12.13741	TAATTAGCAATGGTCACTTT
PMT1061	119	13.60201	ATTGTTACTATTGAAACCAC
PMT1213	147	13.57789	ATTGGGATCTTTGATAAAAA
PMT1244	268	13.1531	TCTGTAGCAACAAAAACCGA
PMT1276	141	16.31724	GATGTGGCTTTTGCTACCTG
PMT1277	28	16.46187	CAGGTAGCAAAAGCCACATC
PMT1480	47	15.52831	TCGGTAACAACAGCCACAAC
PMT1579	172	12.45182	TATGGATGAATAAGCACAGT
PMT1758	72	14.00264	AATGTGCCTGATGCTCCCTG
PMT1831	138	19.22064	ACCGTCACCATTGCTACATG
PMT1853	100	14.32158	TTCGTATCAACATGAACATA
PMT1979	101	13.61964	ATTGAAACCATTGATACATC
PMT1980	278	14.1815	GATGTATCAATGGTTTCAAT
PMT2229	97	12.04587	TTTGTATCATTCACTACAGA
PMT2246	176	19.46608	AAAGTACTCTTTGCTACCTA
Pro0080	120	12.18497	AATGTTTTCAATAATACAAA
Pro0088	157	13.48113	TAAGTACCCTTTTATACCAA
Pro0276	20	22.82149	CTTGTATCAAAAATGACTTT
Pro0277	65	23.04317	AAAGTCATTTTTGATACAAG
Pro0369	279	12.95705	TTGTGATCAAAGATCACATT
Pro0415	99	14.90975	TTTGTAGCGATTGCTACAAA
Pro0416	220	15.23282	TTTGTAGCAATCGCTACAAA
Pro0525	71	12.40095	NTGGTCTCAATGGTTACAGC
Pro0807	230	17.61154	TTGGTATAAATACTTACTTT
Pro0874	270	12.60331	TATGTATAAAAGATAAGAAT
Pro0884	142	12.56437	TTCGTATCATTCGTCACTAT
Pro1037	187	24.75488	TTTGTATCAATAAGTACTTT
Pro1038	71	24.04508	AAAGTACTTATTGATACAAA
Pro1209	280	12.66658	TCTGTATTAACAGATACATA
Pro1262	189	16.20907	TATGTAAGAAAAATCACTTT
Pro1270	61	13.10047	TATGTCTCATCAGTAGCATT
Pro1276	72	12.03402	CAAGTATCAAAAATTACAAG
Pro1466	29	16.23659	TTTGTATCAAATGTCACATG
Pro1471	44	20.03587	CTGGTATCAACACGTACATC
Pro1614	146	19.67509	TTTGTAATTGATGATACATA
Pro1615	67	19.15405	TATGTATCATCAATTACAAA
Pro1620	42	13.45334	TATGAATCACCAATAACTAT
ss1RNA_7	51	12.70874	AGCGTTAGCGTTGATATAAG
SYNW0314	80	12.40029	GATGTAATTGCTGCTCCAGG
SYNW0382	36	16.35403	TAGGTAGCAACGGTTACTGG

SYNW0388	76	15.61574	AACGTCACTCTTGATACAAT
SYNW0389	79	15.69087	ATTGTATCAAGAGTGACGTT
SYNW1073	77	25.35258	AATGTGCGCGTTGATACAAA
SYNW1074	164	25.44243	TTTGTATCAACGCGCACATT
SYNW1351	129	14.62308	TCTGTATCAGCAGCTACCAA
SYNW1352	116	14.63848	TTGGTAGCTGCTGATACAGA
SYNW1357	53	12.39165	TTCGTATCAGGACGAACAAT
SYNW1358	90	12.39574	ATTGTTCGTCCTGATACGAA
SYNW1557	217	14.28763	ACTGGCGCTGTTGTTACCAA
SYNW1564	291	12.15362	ATAGACACTATTGATAGCAA
wh8RNA_24	85	14.98478	CTGGTTTCAACGGTCACAGA
SYNW1633	93	14.97918	TCTGTGACCGTTGAAACCAG
SYNW1951	59	20.23161	TTGGTAGCAGCACTCACAGT
SYNW2255	52	15.65645	CATGTGTCAACGATGACGGT
SYNW2442	69	14.25096	TCGGTTCCGGTTGATACCAA
SYNW2450	150	12.35086	TTGGTATCAAGATGAACAAC
5. motif; matrix 18		(cutoff score: 10)	
PMM0075	93	12.09851	GTAGTCAGAAGATCCTGACG
PMM0075	91	29.45427	AGTCAGAAGATCCTGACGAC
PMM1133	245	10.76102	GTCGCCTGGAACTCCTGATT
PMT0117	54	10.03596	GTTGACAGGGTCGGCCGACC
PMT0170	193	11.48746	GTAGTCATGTTGCTCTGTCT
PMT0193	44	10.67474	GTCAGCAGGAGCATCTGACA
PMT0409	37	13.23403	TTGCAGAAGATGTTGACCAC
PMT0552	176	11.2638	TGTCAGTAGATTTTGATTAA
PMT0553	167	11.21573	TTAATCAAAATCTACTGACA
PMT0618	140	10.82812	GGTCAGTTGGTGCTGTTGCT
PMT0742	237	10.35142	GATAATTTGATCCTGATGAC
PMT0863	96	10.48249	AGTCAACAGGTGCTAGCCAC
PMT0864	248	10.23127	GTGGCTAGCACCTGTTGACT
PMT1081	129	11.24873	AGCATCAGCATCAGCTCACC
PMT1144	35	11.66163	AGTCTGCAGACGCTGACCTG
PMT1220	146	12.41705	AGTCAGTGGATGGTGGGGAG
PMT1277	150	17.22112	NTGGACAAGATCTACTGCCC
PMT1294	21	10.01894	ATGGTCAGATTTTCATGACC
PMT1447	37	10.93535	AGTCAGAGCATGATCAACAC
PMT1565	217	10.83992	AGTCTGAGGACGCAGACACC
PMT1581	227	16.54442	GTGCTCAGGATCCTTTGAGT
PMT1608	97	12.90467	GCGGTCAGTAGATGCTGACC
PMT1608	95	30.54717	GGTCAGTAGATGCTGACCAC
PMT2030	170	16.55756	AGGTTCAGGATCTATTGCCC
mitRNA_25	248	16.53636	GGGCAATAGATCCTGAACCT
Pro0088	98	11.52647	GCAGTCAGTAGATCCTGACT
Pro0088	96	30.5446	AGTCAGTAGATCCTGACTAC
Pro0839	170	12.85559	AGCCAGAATAGCTTGACTAC
Pro0840	58	12.81418	GTAGTCAAGCTATTCTGGCT
SYNW0009	46	10.66969	CTGGTTAGGTTGGTCTGACT
SYNW0238	205	12.0862	GCTCGGTGGCTGCAGACGAC
SYNW0313	72	28.09206	AGGGTCAGGATCCACTGACC
SYNW0314	23	14.72306	CGGCAGTGGATGCTGAAGCT
SYNW0476	137	10.46759	GACGTCAGCAGCTCCGCACC
SYNW0477	32	11.10806	GGTGCGGAGCTGCTGACGTC
SYNW0532	70	13.06778	GTGATCGGCATTCACTGCCC
SYNW0533	52	13.20797	GGGCAGTGAATGCCGATCAC
SYNW0577	182	13.12466	GGTCAGTCCTTGCTGATCCC
SYNW0710	36	15.47353	GGAGTCAGCGTCCGCTGATT
SYNW0888	161	13.72338	GTTCACCAGAGGCTGACGAC
SYNW0947	53	10.07786	GCCGTCTGGATTGTCAGACT
SYNW1139	285	14.00371	GTGGATAGGGTCCACGGACT
SYNW1317	118	11.5402	GTTCAGAGGATCGTGATCGT
SYNW1430	235	10.76317	TTCCAGTGGATCCTGCTCAC
SYNW1441	51	10.43215	TGTCAGCACATGCAGACAAT
SYNW1470	60	11.38806	ATAGTCTTGATCCACCGATC
SYNW1562	53	10.8299	ATTGTCACGATCTGATCACC
SYNW1653	67	10.64802	ATCGTCAGGTGCGGCTGAGC
SYNW1654	29	10.70775	GCTCAGCCGCACCTGACGAT
SYNW1922	24	11.25092	AGTCAGGGGATTCAGAAGCT
SYNW1923	64	11.04157	AGCTTCTGAATCCCCTGACT
SYNW2107	140	10.06795	GTCGTGAGCAGCTCCTGGGC
SYNW2303	254	10.22033	GGTCACAAGAGCCATACCAG
SYNW2336	204	11.08036	AGGTTCAGGCTTTTCAGACC
SYNW2365	34	10.70607	AGTCCGCAGATGTTCACTCT

SYNW2427	251	10.99942	CGATTCAGCATCGTCTGAGT
SYNW2491	33	11.50884	GTTCAGGAGTGCCGGACCCC
SYNW2492	45	11.45128	GGGGTCCGGCACTCCTGAAC

Table 2: List of predictions for putative TF binding sites in marine cyanobacteria, *Prochlorococcus* Med4, MIT 9313, SS120 and *Synechococcus* WH 8102, calculated with the five best PSWMs (based on matrix 2 (ArsR-like), 8, 11 (LexA), 12 (NtcA) and 18 given in Tab. 2), which were obtained by phylogenetic footprinting in these strains.

Strain				Align.	Z	Z rev	Exp	Comment
MED	SS1	MIT	WH8	length				
3	-	-	-	201	-6.28	**-12.94**	+	*yfr2*, *yfr3*, *yfr4*
-	1	1	1	345	-7.58	**-10.18**	n.t.	*rplCD* operon leader, corresponds to *E. coli* S10 r-operon
1	1	1	2	756	-4.47	**-9.9**	n.t.	*rrn* operon leader
2	-	-	-	1129	-8.15	**-9.15**	n.t.	reciprocal coverage of 7.9%, artifact due to low-complexity sequences
-	-	1	1	161	**-7.98**	-4.90	n.t.	highly similar sequences, putative ncRNAs
-	-	1	1	229	**-7.38**	-7.32	n.t.	*rplJ* operon leader, corresponds to *E. coli* β r-operon
-	1	-	-	122	**-6.27**	-5.54	+	*yfr2*
-	1	1	1	152	**-5.77**	-5.61	n.t.	putative Cobalamin riboswitch
-	-	1	1	142	**-5.29**	-5.28	n.t.	possible bidirectional terminator of the *rplKAJL* operon
-	1	1	4	397	-4.38	**-4.95**	n.t.	no conserved position, no significant BLASTN hit to MED4
1	1	-	-	146	-0.84	**-4.92**	+	*yfr7*
2	2	1	4	697	-3.26	**-4.59**	+	*yfr6* in MED4 and SS120 and a subgroup of 5' UTR regions to annotated genes and putative unannotated genes in all 4 strains
-	-	1	1	259	-3.7	**-4.53**	-	*rpoBC* operon leader, corresponds to *E. coli* attenuator separating the *rp*l genes from *rpoBC* in the *rplKAJLrpoBC* gene cluster
1	-	-	-	153	-1.63	**-4.28**	-	Located between genes for a two-component sensor histidine kinase and a cons. hypoth. protein
-	-	1	1	336	**-4.24**	-3.64	n.t.	region upstream of the *rbcLS* cluster containing conserved promoter
-	-	1	1	106	-0.67	**-4.00**	n.t.	*rpl11* operon leader, corresponds to *E. coli* L11 r-operon
-	-	1	1	176	-3.42	**-3.97**	+	*yfr1*
-	1	1	1	197	**-3.93**	-2.94	n.t.	putative TPP riboswitch in front of *thiC*

Figure 3: List of high scoring clusters revealed by RNA prediction in marine cyanobacteria (Axmann et al., 2005). RNA elements were predicted according to the scheme shown in Materials and Methods, Figure 2.3. The total number of sequences in each cluster and the distribution within the four compared genomes plus the total alignment length are given. The elements are ordered according to the lowest score in either forward (Z) or reverse (Z rev) orientation (in bold letters). The lower the Z-score the higher the support for structural conservation. Exp (experimental testing): +, tested positively by Northern hybridisation; NT, not tested. The cluster identities (CLID) were also used in Results, Table 3.5. For further details and exact positions of sequences see Results, Table 3.5 and Kensche (2005); Axmann et al. (2005).

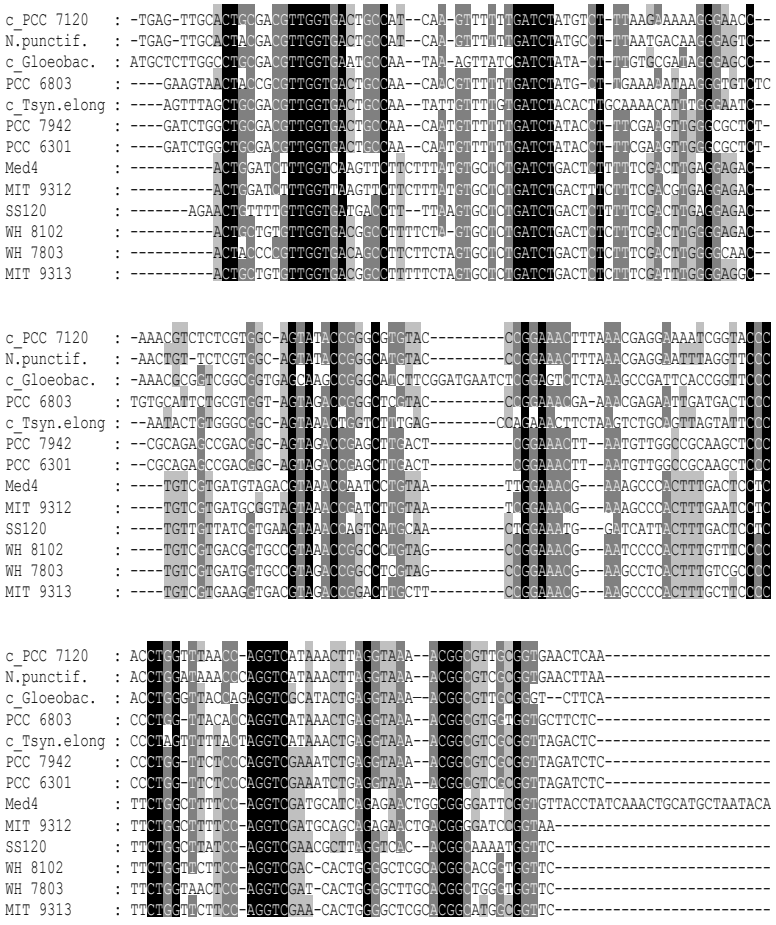


Figure 4: 6Sa alignment of 13 sequences from diverse cyanobacteria from *Anabaena* PCC 7120, *Nostoc punctiforme*, *Synechocystis* PCC 6803, *Gloeobacter violaceus* PCC 7421, *Thermosynechococcus elongatus*, *Synechococcus* PCC 7942, *Synechococcus* PCC 6301, *Prochlorococcus* MIT 9313, *Synechococcus* WH 7803, WH 8102, *Prochlorococcus* SS120, Med4 and MIT 9312.

Abbreviations

A, G, C, T, U	bases of the DNA and RNA, respectively
aa	amino acid
ATCC	American Tissue Culture Collection
ATP	adenosintriphosphate
bp	base pairs
cDNA	complementary DNA
Ci	Curie
DCMU	3-(3,4-dichlorophenyl)-N-N'-dimethylurea
DMSO	dimethyl sulfoxide
DNA	deoxyribonucelic acid
E	Einstein: 6 x 10^{23} photons
EDTA	ethylendiaminetetraacetic acid
EtOH	ethanol
g	constant of gravitation
h	hour
HSPs	high-scoring segment pairs
IGR	intergenic spacer region
kb	kilobase pairs
l	liter
L/D	light-dark
M	mole
Mb	megabase pairs
min	minutes
MIT	Massachusetts Institute of Technology
ml	milliliter
mRNA	messenger RNA
μ	micro (10^{-6})
n	nano (10^{-9})
NaOAc	sodium acetate
ncRNA	non-coding RNA
nt	nucleotide
ORF	open reading frame
p	pico (10^{-12})
PCC	Pasteur Culture Collection
PCR	polymerase chain reaction
pH	potentia hydrogenii
RNA	ribonucleic acid
rRNA	ribosomal RNA
RT	reverse transcription
SDS	sodiumdodecylsulfate
s	seconds
TEMED	N,N,N',N'-Tetramethylethylendiamine
TF	transcription factor
TFBS	transcription factor binding site
TIS	transcriptional initiation site
tRNA	transfer RNA
UTR	untranslated region
UV	ultraviolet
vol/vol	volume per total volume
vol	volume
w/v	weight per volume (g/100ml)
Yfr	cyanobacterial functional RNA

Acknowledgement

Herrn Prof. Thomas Börner darf ich als erstes danken für die Möglichkeit, diese Arbeit am Institut für Biologie der HU Berlin anzufertigen. Damit auch ein ganz besonderer Dank an alle Mitarbeiter der Genetik (vor allem an Caro, Anne R., Reimo, Tobias, Kristina, Conny, Brita, Holger und Elke) - die super Arbeitsatmosphäre und Eure große Hilfsbereitschaft haben viele Experimente erfolgreich verlaufen lassen und die weniger erfolgreichen schnell vergessen gemacht.
Ein ganz großes Dankeschön an Martin - The Lord of the Sequences und aller molekularbiologischen Methoden.
Annegret und Ulf habt Dank für die perfekte Zusammenarbeit und die aufregenden Monate in 2005.

Einen herzlichen Dank ebenso an Jörg - immer bereit sein unerschöpfliches Wissen über RNAs zu teilen.

Vielen, vielen Dank an Szymon, Philip, Nils und Lutz, die nie aufgegeben haben sich meiner Computerfragen anzunehmen; gemeinsam konnten wir einige interessante Informationen aus den ATGC's entschlüsseln.

Many, many thanks to the 'MARGENES clique', we had a nice time together discussing short and complex questions about science and everyday live as well.

Julia, lieben Dank für die Zusammenarbeit und Unterstützung beim Probelesen so vieler Seiten.

I would like to thank Debbie Lindell for the friendly cooperation and the introduction into one of the most fascinating life forms - viruses.

Ein großes Dankeschön an Wolfgang, der wahrscheinlich niemals müde wird neue Ideen zu diskutieren und damit Menschen zu begeistern, stets bereit eine email zu beanworten oder auch mal sonntags zu telefonieren. So waren die letzten drei Jahre ausgefüllt mit spannenden Experimenten (und langen Labortagen), aber auch unterbrochen von äußerst fröhlichen workshops und meetings in den verschiedensten europäischen Ländern.

Natürlich meiner ganzen großen Familie vielen lieben Dank für all die schönen Wochenenden in der Lausitz und im Rheinland zwischen den anstrengenden Arbeitstagen.

Lieben Dank vor allem Iris und Anne, Nina, Jörn und Jan für eure beständige Freundschaft trotz zahllosen Nicht-Anrufen und Absagen.
Den liebsten Dank an Sascha - du bist immer für mich da, und ich bin wegen dir hier, ILD.

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Arbeit selbständig ohne fremde Hilfe verfaßt und nur die angegebene Literatur und Hilfsmittel verwendet zu haben.

Berlin, den 8. März 2006

Ilka Axmann